Multi-resolution localization of causal variants across the genome

Matteo Sesia, Eugene Katsevich, Stephen Bates, Emmanuel Candès*, Chiara Sabatti*

Stanford University, Department of Statistics, Stanford, CA 94305, USA

Abstract

We present *KnockoffZoom*, a flexible method for the genetic mapping of complex traits at multiple resolutions. *KnockoffZoom* localizes causal variants precisely and provably controls the false discovery rate using artificial genotypes as negative controls. Our method is equally valid for quantitative and binary phenotypes, making no assumptions about their genetic architectures. Instead, we rely on well-established genetic models of linkage disequilibrium. We demonstrate that our method can detect more associations than mixed effects models and achieve fine-mapping precision, at comparable computational cost. Lastly, we apply *KnockoffZoom* to data from 350k subjects in the UK Biobank and report many new findings.

* Corresponding authors

I. INTRODUCTION

Since the sequencing of the human genome, there have been massive efforts to identify genetic variants that affect phenotypes of medical relevance. In particular, single-nucleotide polymorphisms (SNPs) have been genotyped in large cohorts in the context of genome-wide association studies (GWAS), leading to the discovery of thousands of associations with different traits and diseases.[1] However, it has been challenging to translate these findings into actionable knowledge.[2] As a first step in this direction, we present a new statistical method for the genetic mapping of complex phenotypes that improves our ability to resolve the location of causal variants.

The analysis of GWAS data initially proceeded SNP-by-SNP, testing *marginal* independence with the trait (not accounting for the rest of the genome) using univariate regression. Today the leading solutions rely on a linear mixed model (LMM) that differs from the previous method insofar as it includes a random term approximating the effects of other variants in distant loci.[3–8] Nonetheless, one can still think of the LMM as testing marginal independence because it does not account for linkage disequilibrium[9,10] (LD). Thus, it cannot distinguish causal SNPs from nearby variants that may have no interesting biological function, since neither are independent of the phenotype. This limitation becomes concerning as we increasingly focus on polygenic traits and rely on large samples; in this setting it has been observed that most of the genome is correlated with the phenotype, even though only a fraction of the variants may be important.[11] Therefore, the null hypotheses of no association should be rejected for most SNPs,[2] which is a rather uninformative conclusion. The awareness among geneticists that marginal testing is insufficient has led to the heuristic practice of post-hoc aggregation (*clumping*) of associated loci in LD[8,12] and to the development of fine-mapping.[13] Fine-mapping methods are designed to refine marginally significant loci and discard associated but non-causal SNPs by accounting for LD, often within a Bayesian perspective.[14–17] However, this two-step approach is not fully satisfactory because it requires switching models and assumptions in the middle of the analysis, obfuscating the interpretation of the findings and possibly invalidating type-I error guarantees. Moreover, as LD makes more non-causal variants appear marginally associated with the trait in larger samples, the standard fine-mapping tools face an increasingly complicated task refining wider regions.

KnockoffZoom simultaneously addresses the current difficulties in locus discovery and fine-mapping by searching for causal variants over the entire genome and reporting those SNPs (or groups thereof) that appear to have a distinct influence on the trait while accounting for the effects of all others. This search is carried out at multiple resolutions, which determine the width of the reported groups, ranging from that of locus discovery to that of fine-mapping. In this process, we precisely control the false discovery rate (FDR)[18] and obtain findings that are directly interpretable, without heuristic post-processing. Our inferences rest on the assumption that LD is adequately described by hidden Markov models (HMMs) that have been successfully employed in many areas of genetics.[19–22] We do not require any model linking genotypes to phenotypes; therefore, our approach seamlessly applies to both quantitative and qualitative traits.

This work is facilitated by recent advances in statistics, notably *knockoffs*,[23] whose general validity for GWAS has been explored and discussed before.[24–29] We leverage these results and introduce several key innovations. First, we develop new algorithms to analyze the data at multiple levels of resolution, in such a way as to maximize power in the presence of LD without pruning the variants.[23,24] Second, we improve the computational efficiency and apply our method to a large dataset, the UK Biobank[30]—a previously computationally unfeasible task.

II. RESULTS

A. *KnockoffZoom*

To localize causal variants as precisely as possible and provide researchers with interpretable results, *KnockoffZoom* tests *conditional hypotheses* (Online Methods A). In the simplest GWAS analysis, a variant is null if the distribution of its alleles is independent of the phenotype. By contrast, our hypotheses are much stricter: a variant is null if it is independent of the trait *conditionally* on all other variants, including its neighbors. Suppose that we believed, consistent with the classical polygenic literature,[31] that a multivariate linear model realistically describes the genetic makeup of the trait; then, a conditional hypothesis is non-null if the corresponding variant has a non-zero coefficient. In general, in the absence of unmeasured confounders or any feedback effects of the phenotype onto the genome, our tests lead to the discovery of causal variants. In particular, we can separate markers that have a distinct effect on the phenotype from those whose associations are merely due to LD.

The presence of LD makes the conditional hypotheses challenging to reject, especially with small sample sizes. This is why t-tests for multivariate linear regression may have little power in the presence of collinearity; in this case, F-tests provide a possible solution. Analogously, we group variants that are too similar for their distinct effects to be discerned. More precisely, we test whether the trait is independent of all SNPs in an LD block, conditional on the others.

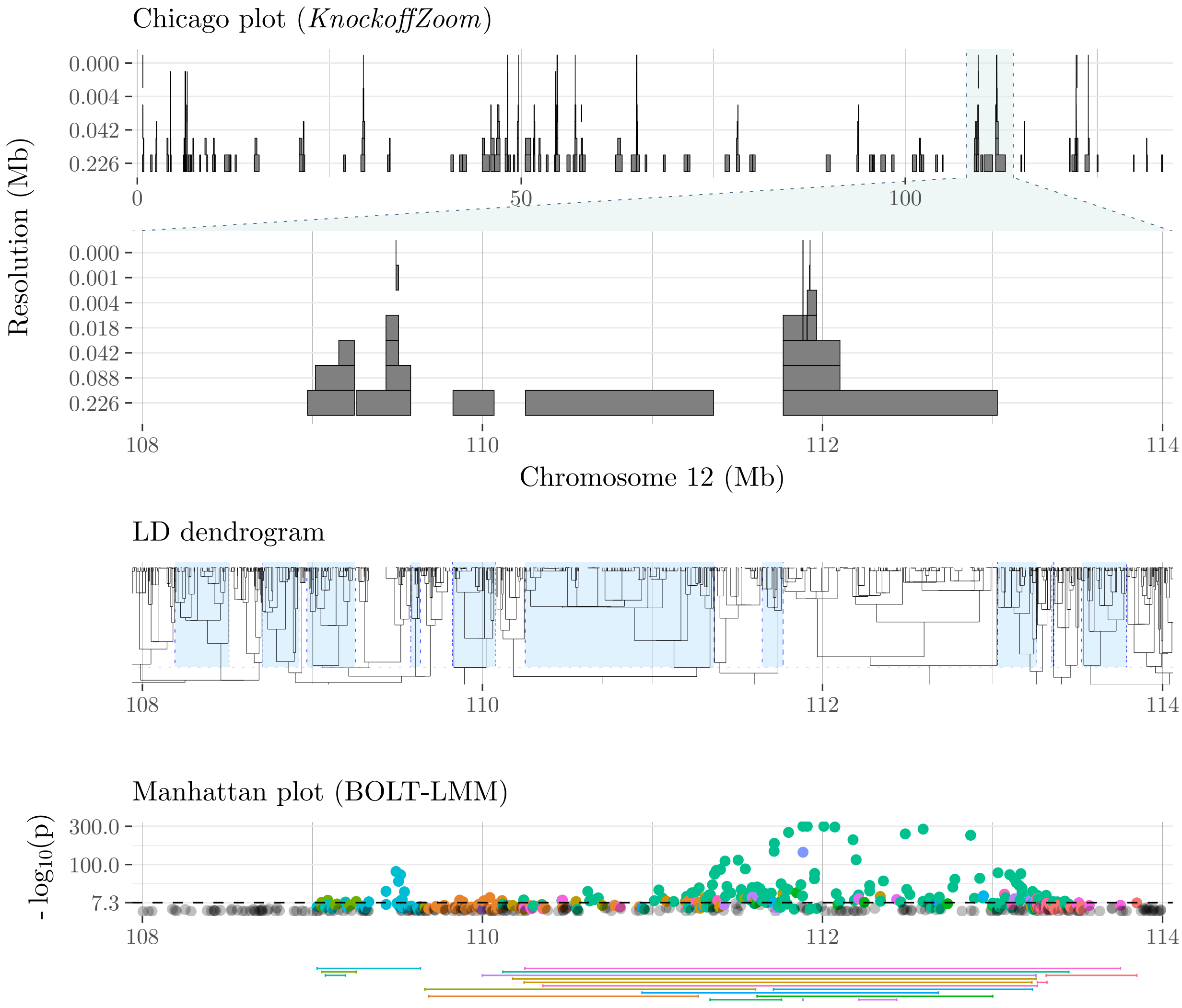


FIG. 1. (a): *KnockoffZoom* discoveries on chromosome 12 for the phenotype *platelet* in the UK Biobank, controlling the FDR at the nominal level 0.1. Each shaded rectangle represents a discovery made at the resolution (measured by the average width of the blocks) indicated by its vertical position, so that the highest-resolution findings are on top. The lower part of (a) focuses on a smaller segment of the chromosome. The hypotheses are pre-specified by cutting at different heights the LD dendrogram sketched in (b). For example, by alternating light blue and white shading in this dendrogram, we indicate the blocks corresponding to the lowest resolution. The Manhattan plot in (c) shows the BOLT-LMM p-values computed from the same data, while the segments below represent the region spanned by the significant discoveries clumped with PLINK at the genome-wide significance level (5×10^{-8}). For each clump, the colors match those of the corresponding p-values. All plots are vertically aligned, except for the upper part of (a).

A block is null if it only contains SNPs that are independent of the phenotype, conditional on the other blocks. Concretely, we define the blocks at multiple resolutions by partitioning the genome via adjacency-constrained hierarchical clustering.[32] We adopt the r^2 coefficients computed by PLINK[12] as a similarity measure for the SNPs and cut the dendrogram at different heights; see Figure 1 (b). Alternative grouping strategies could be easily integrated into *KnockoffZoom*. To balance power and resolution, we consider multiple partitions, starting with a coarse view and successively refining it (Supplement, Table S3). An example of our results for *platelet* is shown in Figure 1 (a). Each rectangle in the *Chicago* plot (named for its stylistic resemblance to the Willis tower) spans a genomic region that includes variants carrying distinct information about the trait compared to the other blocks on the same level. The blocks at higher levels correspond to higher resolutions; these are narrower and more difficult to discover, since only strong associations can be refined precisely.

We can mathematically prove that *KnockoffZoom* controls the FDR below any desired level q (at most a fraction q of our findings are false positives on average), at each resolution. The FDR is a meaningful error rate for the study of complex traits,[33,34] but its control is challenging.[35] We overcome this difficulty with *knockoffs*. These are carefully engineered synthetic variables that can be used as negative controls because they are exchangeable with the genotypes and reproduce the spurious associations that we want to winnow.[23,24,36,37] The construction of knockoffs requires us to specify the distribution of the genotypes, which we approximate as a HMM. In this paper, we implement the HMM of fastPHASE,[19,20] although our framework is sufficiently flexible to accommodate other choices in the future. Then, we extend an earlier Monte Carlo algorithm for generating knockoffs[24] to target the multi-resolution hypotheses defined above (Supplement, Section S1 A). Moreover, we reduce the computational complexity of this operation through exact analytical derivations (Supplement, Section S2). With the UK Biobank data, for which the haplotypes have been phased,[22] we accelerate the algorithm further by avoiding implicit re-phasing (Supplement, Section S2 B 2).

To powerfully separate interesting signals from spurious associations, we fit a multivariate predictive model of the trait and compute feature importance measures for the original and synthetic genotypes (Online Methods B). The contrast between the importance of the genotypes and their knockoffs is used to compute a test statistic in each of the LD blocks defined above, for which a significance threshold is calibrated by the *knockoff filter*.[36] As a predictive model, we adopt an efficient implementation of sparse linear and logistic regression designed for massive genetic datasets,[38] although *KnockoffZoom* could easily incorporate other methods.[23] Thus, *KnockoffZoom* exploits the statistical power of variable selection techniques that would otherwise offer no type-I error guarantees.[23,31] A full schematic of our workflow is in the Supplement (Figure S1), while software and tutorials are available from `https://msesia.github.io/knockoffzoom/`. The computational cost (Supplement, Table S1) compares favorably to that of alternatives, e.g., BOLT-LMM.[7,8]

Revisiting Figure 1, we note that our discoveries are clearly interpretable because they are distinct by construction and each suggests the presence of an interesting SNP. The interpretations of hypotheses at different resolutions are easily reconciled because our partitions are nested (each block is contained in exactly one larger block from the resolution below), while the null hypothesis for a group of variants is true if and only if all of its subgroups are null.[39] Most of our findings are confirmed by those at lower resolution even though this is not explicitly enforced and some "floating" blocks are occasionally reported, as also visible in Figure 1. These are either false positives or true discoveries that have not reached the adaptive significance threshold for FDR control at lower resolution. A slight variation of our procedure can explicitly avoid "floating" blocks by coordinating discoveries at multiple resolutions, although with a possible power reduction (Supplement, Section S1 B).

Our findings lend themselves well to cross-referencing with gene locations and functional annotations. An example is shown in Figure 5. By contrast, the output of BOLT-LMM in Figure 1 (c) is less informative: many of the clumps reported by the standard PLINK algorithm (Section II C 2) are difficult to interpret because they are wide and overlapping. This point is clearer in a numerical simulation in which we know the causal variants, as previewed in Figure 2. Here, we see two consequences of stronger signals: our method precisely identifies the causal variants, while the LMM reports wider and increasingly contaminated sets of associated SNPs. The details of our experiments are discussed in Section II C.

B. Conditional hypotheses and population structure

The *KnockoffZoom* discoveries, by accounting for LD, bring us closer to the identification of functional variants; therefore, they are more interesting than simple marginal association. Moreover, conditional hypotheses are less susceptible to the confounding effect of population structure.[29] As an illustration, consider studying the genetic determinants of blood cholesterol levels, blindly sampling from European and Asian populations, which differ by likelihood of the trait (due to diet or genetic reasons) and in the prevalence of several markers (due to different ancestries). It is well known that marginally significant SNPs do not necessarily indicate a biological effect on lipid levels, as they may simply reflect a diversity in the allele frequencies across the populations (which is correlated with the trait). Multivariate association methods, like *KnockoffZoom*, are less susceptible to this confounding effect than single-marker techniques because they account for the

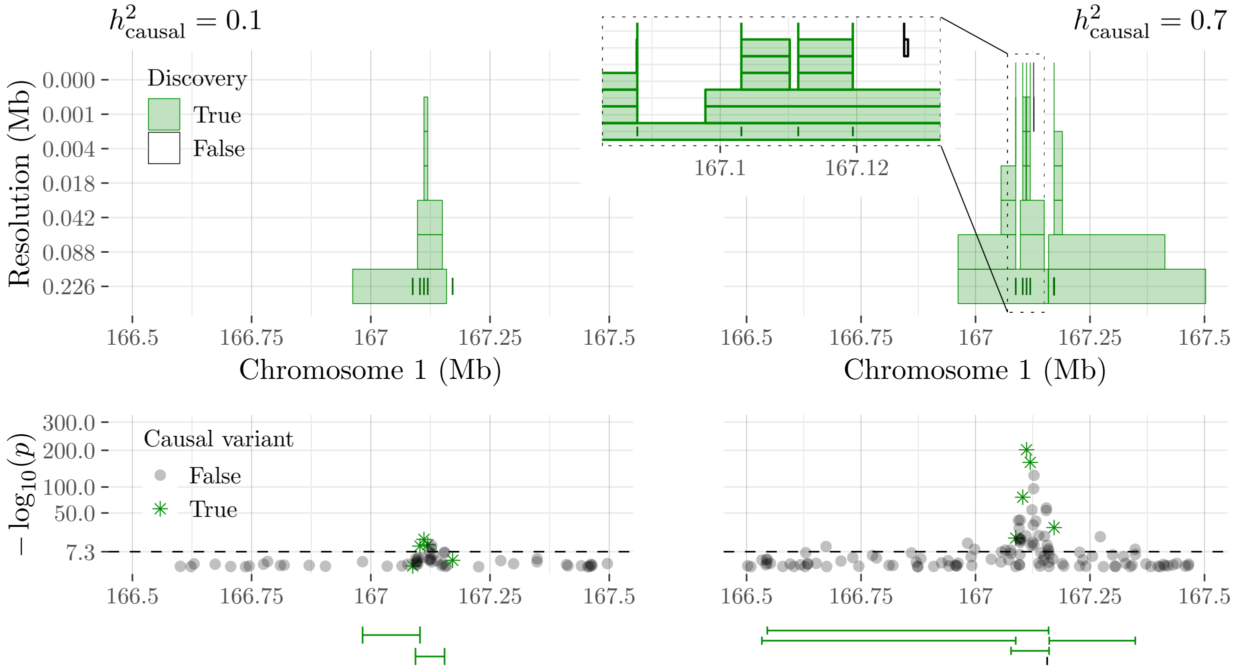


FIG. 2. *KnockoffZoom* discoveries for two simulated traits with the same genetic architecture but different heritability h^2_{causal}, within a genomic window. Other details are as in Figure 1. This region contains 5 causal SNPs whose positions are marked by the short vertical segments at the lowest resolution. The zoomed-in view (right) shows the correct localization of the causal SNPs, as well as a "floating" false positive.

information encompassed in all genotypes, which includes the population structure.[4,5,40]

Conditional hypotheses are most robust at the highest resolution, where we condition on all SNPs except one, capturing most of the genetic variation. The susceptibility to the confounding effect may increase as more variants are grouped and removed from the conditioning set, since these may be informative regarding the population structure. However, we consider fairly small LD blocks, typically containing less than 0.01% of the SNPs, even at the lowest resolution. By comparison, the LMM p-values may not account for the effect of an entire chromosome at once, to avoid "proximal contamination".[7]

Our inferences rely on a model for the distribution of the genotypes, which requires some approximation. The HMM implemented here is more suitable to describe homogeneous and unrelated individuals because it does not capture long-range dependencies, which is a technical limitation we plan to lift in the future. Meanwhile, we analyze unrelated British individuals. Our results already explicitly account for any available covariates, including those that reflect the population structure, such as the top principal components of the genetic matrix;[41] see Section II D. Moreover, we can account for any known structure (labeled populations) with the current implementation of our procedure, by fitting separate HMMs and generating knockoffs independently for each population.

C. Power, resolution and false positive control in simulations

1. Setup

To test and compare *KnockoffZoom* with state-of-the-art methods, we first rely on simulations. These are designed using 591k SNPs from 350k unrelated British individuals in the UK Biobank (Online Methods C). We simulate traits using a linear model with Gaussian errors: 2500 causal variants are placed on chromosomes 1–22, clustered in evenly-spaced 0.1 Mb-wide groups of 5. The effect sizes are heterogeneous, with relative values chosen uniformly at random across clusters so that the ratio between the smallest and the largest is 1/19. The heritability is varied as a control parameter. The causal variables in the model are standardized, so rarer variants have stronger effects. The direction of the effects is randomly fixed within each cluster, using the major allele as reference. This architecture is likely too simple to be realistic,[11] but it facilitates the comparison with other tools, whose assumptions are reasonable in this setting. By contrast, we are not protecting *KnockoffZoom* from model misspecification because we use real genotypes and approximate their distribution (our sole assumption is that we can do this accurately) with the same HMM (Supplement, Section S4 A) as in our data analysis. Therefore, we give *KnockoffZoom* no advantage. We discuss simulations under alternative architectures in the Supplement (Section S4).

In this setup, the tasks of *locus discovery* and *fine-mapping* are clearly differentiated. The goal of the former is to detect broad genomic regions that contain interesting signals; also, scientists may be interested in counting how many distinct associations have been found. The goal of the latter is to identify the causal variants precisely. Here, there are 500 well-separated regions and 2500 signals to be found. For locus discovery, we compare *KnockoffZoom* at low resolution to BOLT-LMM[8]. For fine-mapping, we apply *KnockoffZoom* at 7 levels of resolution (Supplement, Section S4 B) and compare it to a two-step procedure: BOLT-LMM followed by CAVIAR[14] or SUSIE.[17] For each method, we report the power, false discovery proportion (FDP) and mean size of the discoveries. Since locus discovery and fine-mapping have different goals, we need to clearly define false positives and count the findings appropriately. Our choices are detailed below.

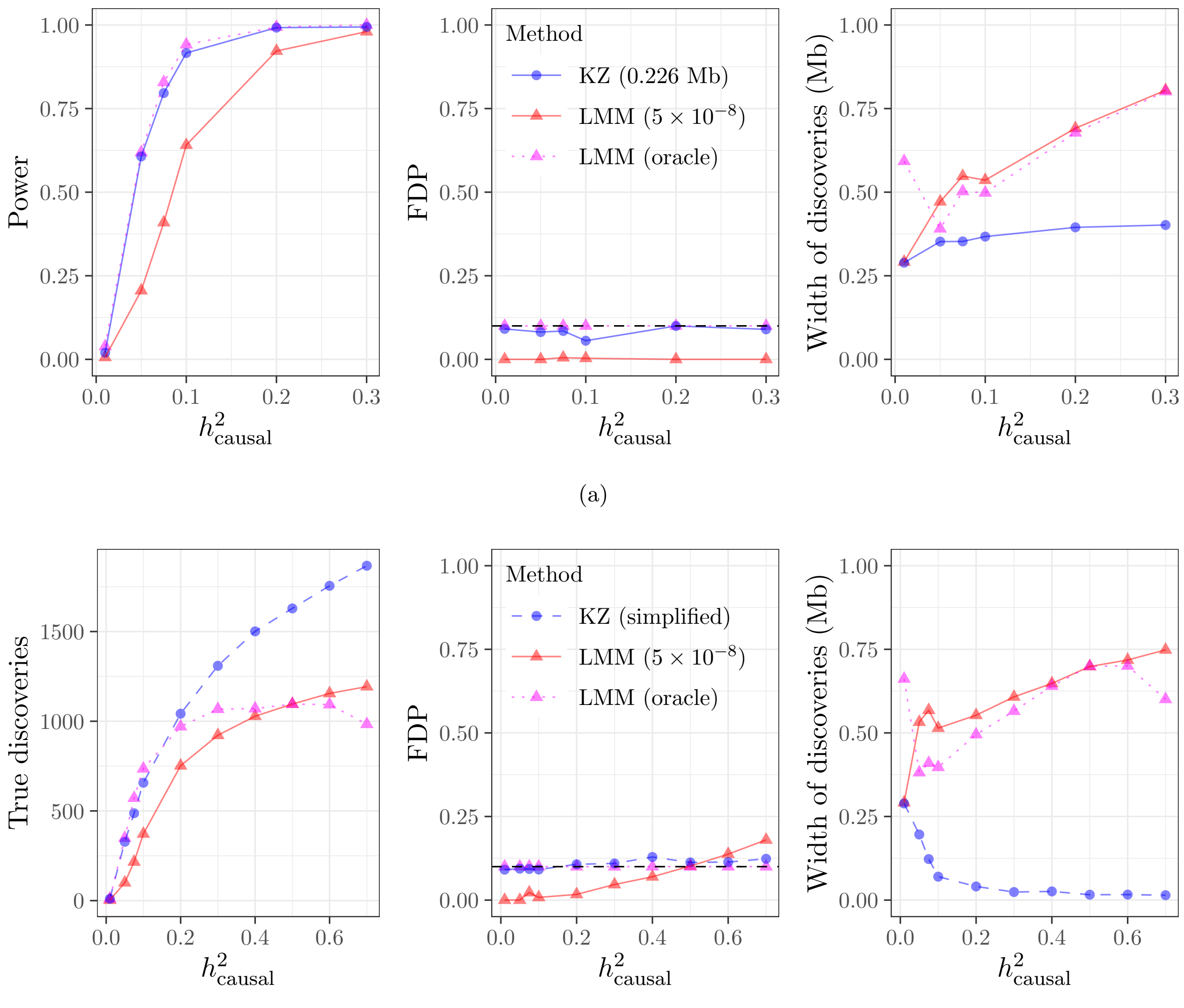
2. *Locus discovery*

We apply *KnockoffZoom* at low resolution, with typically 0.226 Mb-wide LD blocks, targeting an FDR of 0.1. For BOLT-LMM,[8] we use the standard p-value threshold of 5×10^{-8}, which has been motivated by a Bonferroni correction to control the family-wise error rate (FWER) below 0.05, for the hypotheses of no marginal association.

To assess power, any of the 500 interesting genomic regions is detected if there is at least one finding within 0.1 Mb of a causal SNP. This choice favors BOLT-LMM, which is not designed to precisely localize causal variants and reports wider findings (Figure 3). To evaluate the FDP, we need to count true and false discoveries. This is easy with *KnockoffZoom* because its findings are distinct and count as false positives if causal variants are not included. In comparison, distinct LMM discoveries are difficult to define because the p-values test marginal hypotheses. A common solution is to group nearby significant loci. Throughout this paper, we clump with the PLINK[12] greedy algorithm using a 5 Mb window and an r^2 threshold equal to 0.01, while the secondary significance level is 10^{-2} (to include sub-threshold SNPs in LD with an existing clump), as in earlier work.[8] For reference, this is how the clumps in Figure 1 are obtained. Then, we define the FDP as the fraction of clumps whose range does not cover a causal SNP. For locus discovery, we consolidate clumps within 0.1 Mb.

KnockoffZoom and BOLT-LMM target different error rates, complicating the comparison. Ideally, we would like to control the FDR of the distinct LMM findings. Unfortunately, this is difficult[35] (Supplement, Section S4 D). We address this obstacle, within simulations, by considering a hypothetical *oracle* procedure based on the LMM p-values that is guaranteed to control the FDR. The oracle knows in advance which SNPs are causal and uses this information to identify the most liberal p-value threshold such that the fraction of falsely reported clumps is below 0.1. Clearly, this cannot be implemented in practice. However, the power of this oracle provides an informative upper bound for any future concrete FDR-controlling procedure based on BOLT-LMM p-values.

Figure 3 (a) compares the performances of these methods as a function of the heritability. The results refer to an independent realization of the simulated phenotypes for each heritability value. The average behavior across repeated experiments is discussed in the Supplement (Section S4 G), on smaller simulated datasets. The FDP of *KnockoffZoom* is below the nominal level (the guarantee is that the FDP is controlled on average), while its power is comparable to that of the oracle. Moreover, our method reports more precise (narrower) discoveries. The realistic implementation of BOLT-LMM with the standard p-value threshold of 5×10^{-8} is substantially less powerful.



(a)

(b)

FIG. 3. Locus discovery for a simulated trait with *KnockoffZoom* (nominal FDR 0.1) and BOLT-LMM (5×10^{-8} and *oracle*). (a): low-resolution *KnockoffZoom* and strongly clumped LMM p-values. (b): multi-resolution *KnockoffZoom* (simplified count) and weakly clumped LMM p-values.

Before considering alternative fine-mapping methods, we compare *KnockoffZoom* and BOLT-LMM at higher resolutions. Above, the block sizes in *KnockoffZoom* and the aggressive LMM clumping strategy are informed by the known location of the regions containing causal variants, to ensure that each is reported at most once. However, this information is generally unavailable and the scientists must determine which discoveries are important. Therefore, we set out to find as many distinct associations as possible, applying *KnockoffZoom* at multiple resolutions (Supplement, Section S4 B), and interpreting the LMM discoveries assembled by PLINK as distinct, without additional consolidation. An example of the outcomes is visualized in Figure 2. In this context, we follow a stricter definition of type-I errors: a finding is a true positive if and only if it reports a set of SNPs that includes a causal variant. We measure power as the number of true discoveries. Instead of showing the results of *KnockoffZoom* at each resolution (Supplement, Figure S3), we count only the most specific findings whenever the same locus is detected with different levels of granularity, and discard finer discoveries that are unsupported at lower resolutions. This operation is unnecessary to interpret our results and is not sustained by theoretical guarantees,[42] although it is quite natural and informative.

Figure 3 (b) summarizes these results as a function of the heritability. *KnockoffZoom* reports increasingly precise discoveries as the signals grow stronger. By contrast, the ability of BOLT-LMM to resolve different signals does not improve. Instead, stronger signals make more non-causal variants marginally significant through LD, obfuscating the interpretation of the discoveries and making FDR control even more challenging. This cautions one against placing too much confidence in the estimated number of distinct findings obtained with BOLT-LMM via clumping.[8]

3. *Fine-mapping*

We compare *KnockoffZoom* with two fine-mapping methods, CAVIAR[14] and SUSIE,[17] which we apply within a two-step procedure because they are not designed to operate genome-wide. We start with BOLT-LMM (5×10^{-8} significance) for locus discovery and then separately fine-map each reported region. We aggressively clump the LMM findings to ensure that they are distinct, as in Figure 3 (a). We also provide unreported nearby SNPs as input to SUSIE, to attenuate the selection bias (Supplement, Figure S4). Within each region, CAVIAR and SUSIE report sets of SNPs that are likely to contain causal variants. We tune their parameters to obtain a genome-wide FDR comparable to our target (Supplement, Section S4 F).

The results are shown in Figure 4, defining the power and FDP as in Figure 3 (b). The output of *KnockoffZoom* is presented in two ways: at a fixed resolution (e.g., 0.042 Mb; see Supplement, Table S3) and summarizing the multi-resolution results as in Figure 3 (b). All methods appear to control the FDR and detect the 500 interesting regions as soon as the heritability is sufficiently large, as seen earlier in Figure 3. Moreover, they report precise discoveries, each including only a few SNPs. CAVIAR cannot report more than 500 discoveries here, due to its intrinsic limitations discussed above. SUSIE and *KnockoffZoom* successfully identify more distinct signals, with comparable performance in this setting. A more detailed analysis highlighting their differences is available in the Supplement (Section S4 F).

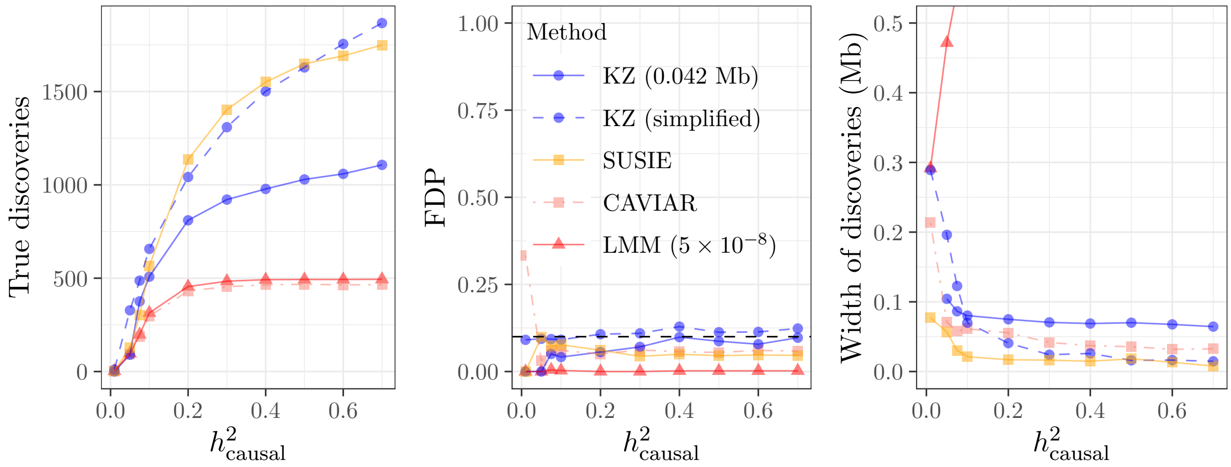


FIG. 4. *KnockoffZoom* compared to a two-step fine-mapping procedure consisting of BOLT-LMM followed by CAVIAR or SUSIE, in the same simulations as in Figure 3. Our method, CAVIAR and SUSIE control a similar notion of FDR at the nominal level 0.1.

D. Analysis of different traits in the UK Biobank data

1. *Findings*

We apply *KnockoffZoom* to 4 continuous traits and 5 diseases in the UK Biobank (Supplement, Table S5), using the same SNPs from 350k unrelated individuals of British ancestry (Online Methods C) and the setup of Section II C. Our discoveries are compared to the BOLT-LMM findings in Table I. We apply our method accounting for the top 5 principal components[41] and a few other covariates in the predictive model (Online Methods D). The results do not change much if we ignore the principal components (Supplement, Table S6), confirming the robustness to population structure. *KnockoffZoom* misses very few of the BOLT-LMM discoveries (see Online Methods B for more details about the case of *glaucoma*) and reports many additional findings (Supplement, Table S7), even when the latter is applied to a larger sample[8] (Supplement, Table S8). The interpretation of the LMM findings is unclear because many clumps computed by PLINK are overlapping, as shown in Figure 1. As we consolidate them, they become distinct but they decrease in numbers, inconsistently with the results reported by others using the same data.[8]

As the resolution increases, we report fewer findings, which is unsurprising since high-resolution conditional hypotheses are more challenging to reject. Our multi-resolution findings can be compared to the known gene positions and the available functional annotations (based on ChromHMM,[43] GM12878 cell line) to shed more light onto the causal variants. We provide an online tool to explore these results interactively (`https://msesia.github.io/knockoffzoom/ukbiobank`); for examples, see Figure 5 and Supplement, Section S5 G.

Phenotype	*KnockoffZoom* resolution (Mb)							BOLT-LMM	
	0.226	0.088	0.042	0.018	0.004	0.001	0.000	clumped	clumped and consolidated
height	3284	1976	823	388	336	170	173	1685	795
bmi	1804	555	60	33	24	0	15	389	328
platelet	1460	890	408	276	161	181	143	723	428
sbp	722	297	95	0	0	0	0	197	178
cvd	514	182	51	0	0	0	0	156	136
hypothyroidism	212	108	0	0	0	0	21	96	77
respiratory	176	65	41	13	14	12	0	63	47
diabetes	50	33	21	10	11	10	0	47	42
glaucoma	0	0	0	0	0	0	0	5	5

TABLE I. Numbers of distinct findings made by *KnockoffZoom* at different resolutions (FDR 0.1), on some phenotypes in the UK Biobank, compared to BOLT-LMM (p-values $\leq 5 \times 10^{-8}$). Data from 350k unrelated British individuals. The LMM discoveries are counted with two clumping heuristics, as in Figure 3.

2. *Reproducibility*

To investigate the reproducibility of the low-resolution findings obtained with different methods on real data, we study *height* and *platelet* on a subset of 30k unrelated British individuals and verify that the discoveries are consistent with those previously reported for BOLT-LMM applied to all 459k European subjects.[8] For this purpose, a discovery is replicated if it falls within 0.1 Mb of a SNP with p-value below 5×10^{-9} on the larger dataset. We do not consider the other phenotypes because, with this sample size, both methods make fewer than 10 discoveries for each of them. The LMM findings are clumped by PLINK without consolidation, although this makes little difference because extensive overlap occurs only with larger samples. To illustrate the difficulty of controlling the FDR with the LMM (Supplement, Section S4), we try to naïvely apply a pre-clumping Benjamini-Hochberg (BH) correction[18] to the LMM p-values.

The results are summarized in Table II. The FDP of *KnockoffZoom* is below the target FDR. All genome-wide significant BOLT-LMM discoveries in the smaller dataset (5×10^{-8}) are replicated, at the cost of lower power. The BH procedure with the LMM p-values does not control the FDR.

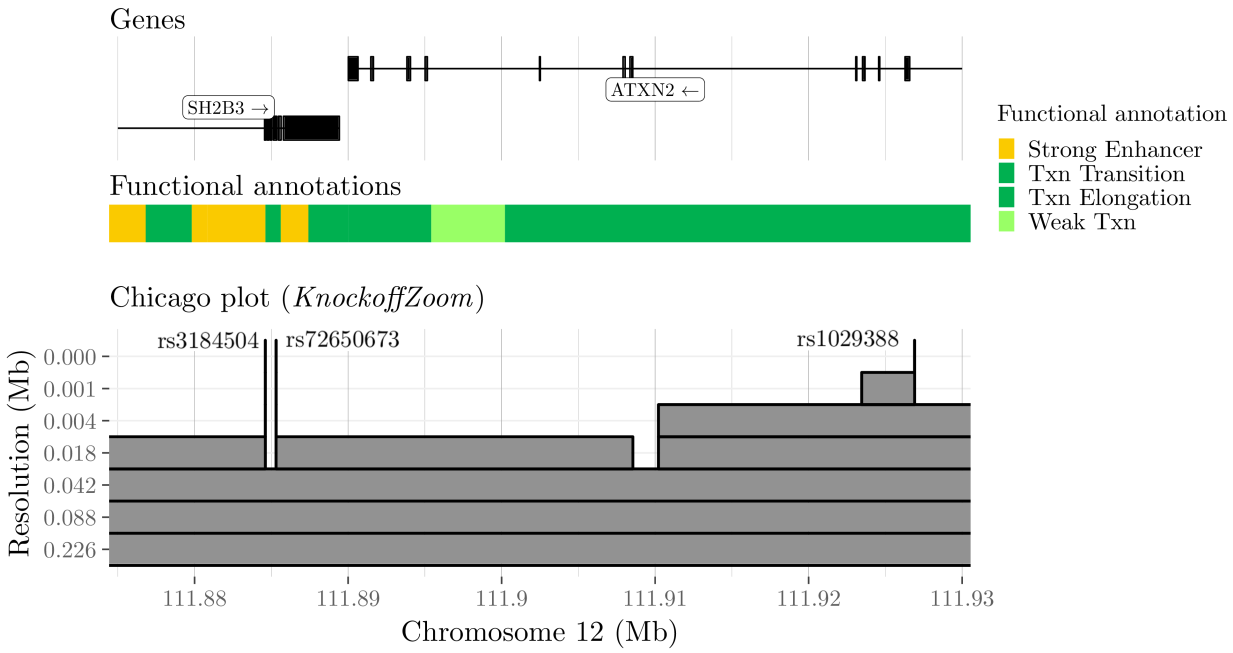


FIG. 5. Visualization of some discoveries made with *KnockoffZoom* for *platelet* in the UK Biobank, along with gene positions and functional annotations (in the LocusZoom[44] style). The three discoveries at single-variant resolution are labeled. Other details as in Figure 1.

The relative power of these alternative methods is discussed in the Supplement (Section S5 D).

		Discoveries		
Phenotype	Method	#	Not replicated	Size (Mb)
height	*KnockoffZoom* (FDR 0.1)	121	8 (6.6%)	0.308
	LMM (5×10^{-8})	54	0 (0.0%)	0.965
	LMM-BH (FDR 0.1)	714	203 (28.4%)	0.379
platelet count	*KnockoffZoom* (FDR 0.1)	81	5 (6.2%)	0.319
	LMM (5×10^{-8})	47	0 (0.0%)	0.674
	LMM-BH (FDR 0.1)	272	92 (33.8%)	0.433

TABLE II. Reproducibility of the low-resolution discoveries made with *KnockoffZoom* (0.226 Mb) and BOLT-LMM on *height* and *platelet*, using 30k individuals in the UK Biobank.

As a preliminary verification of the biological validity of our findings, we use GREAT[45] to conduct a gene ontology enrichment analysis of the regions reported by *KnockoffZoom* for *platelet* at each resolution. The enrichment among 5 relevant terms is highly significant (Supplement, Table S11) and tends to strengthen at higher resolutions, which suggests increasingly precise localization of causal variants.

Finally, we cross-reference with the existing literature the discoveries reported by *KnockoffZoom* at the highest resolution for *platelet*. Note that three of these findings are shown in Figure 5: rs3184504 (missense, SH2B3 gene), rs72650673 (missense, SH2B3 gene) and rs1029388 (intron, ATXN2 gene). Many of the other discoveries are in coding regions and may plausibly localize a direct functional effect on the trait (Supplement, Table S12). Some have been previously observed to be associated with *platelet* (Supplement, Table S13), while others are likely to indicate new findings. In particular, six of our discoveries are missense variants that had not been previously reported to be associated with *platelet* (Supplement, Table S14).

III. DISCUSSION

The goal of genetic mapping is to localize, as precisely as possible, variants that influence a trait. Geneticists have widely sought this goal with a two-step strategy, partly due to computational limitations, in an attempt to control type-I errors while achieving high power. First, all variants are probed in order to identify promising regions, without accounting for LD. To reduce false positives, a Bonferroni correction is applied to the marginal p-values. Then, the roles of different variants are explored with multivariate models, separately within each associated region. This strategy is suboptimal. Indeed, if the phenotypes are influenced by hundreds or thousands of genetic variants, the notion of FWER in the first step is too stringent and inhibits power unnecessarily. This error rate is a legacy of the earlier studies of Mendelian diseases and it has been retained for the mapping of complex traits mostly due to methodological difficulties, rather than a true need to avoid any false positives. In fact, we have shown that the traditional two-step paradigm of locus discovery followed by fine-mapping already effectively tries to control an error rate comparable to the FDR that we directly target. Besides, the type-I error guarantee in the first step is only valid for the SNP-by-SNP marginal hypotheses of no association. These are of little interest to practitioners, who typically interpret the findings as if they contained causal variants. By contrast, *KnockoffZoom* unifies locus discovery and fine-mapping into a coherent statistical framework, so that the findings are immediately interpretable and equipped with solid guarantees.

For ease of comparison, we have simulated phenotypes using a linear model with Gaussian errors, which satisfies many of the assumptions in the standard tests. Unfortunately, there is very little information on how real traits depend on the genetic variants; after all, the goal of a GWAS is to discover this. Therefore, relying on these assumptions can be misleading. In contrast, *KnockoffZoom* only relies on knowledge of the genotype distribution, which we can estimate accurately due to the availability of genotypes from millions of individuals. Indeed, geneticists have developed phenomenological HMMs of LD that work well in practice for phasing and imputation.

It may be useful to highlight with an example that our framework is not tied to any model for the dependence of phenotypes on genotypes or to a specific data analysis tool. We simulate an imbalanced case-control study that it is well-known to cause BOLT-LMM to fail,[46] while our method applies seamlessly. The trait is generated from a liability threshold model (probit), obtaining 525 cases and 349,594 controls. We apply *KnockoffZoom* using sparse logistic regression and report the low-resolution results in Figure 6 as a function of the heritability of the latent Gaussian variable.

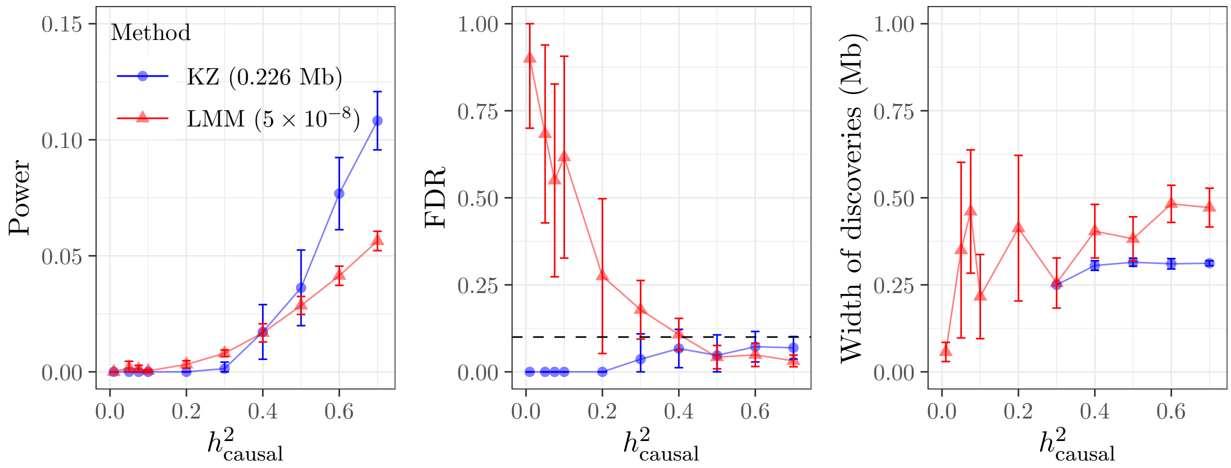


FIG. 6. Locus discovery with *KnockoffZoom* and BOLT-LMM for a simulated case-control study. There are 500 evenly spaced causal variants. The error bars indicate 95% confidence intervals for the mean power and FDR, as defined in Figure 3 (a), averaging 10 independent replications of the trait given the same genotypes.

The released software (`https://msesia.github.io/knockoffzoom/`) has a modular structure that accommodates many diverse options, reflecting the flexibility of our approach. The user may experiment with different importance measures in addition to sparse linear and logistic regression. For example, one can incorporate prior information, such as summary statistics from other studies. Similarly, there are many ways of defining the LD blocks for our analysis: leveraging genomic annotations is a promising direction. Additionally, more refined knockoff constructions may be developed. Indeed, we are working on new algorithms for heterogeneous populations, either relying on an HMM similar to those described for admixture reconstruction,[47,48] or adopting the framework of SHAPEIT.[22] Similarly, it is possible construct knockoffs for related individuals, allowing the analysis of familial data. On a different note, a local measure of the FDR[49] can be used to further quantify the statistical significance of each discovery (Supplement, Sections S4 H and S5 F).

Finally, we highlight that our approach to genetic mapping may naturally lead to the definition of more principled polygenic risk scores, since we already combine multi-marker predictive modeling with interpretable variable selection. Partly with this goal, multi-marker methods for the analysis of GWAS data have been suggested long before our contribution.[31,40,50–52] However, knockoffs finally allow us to obtain reproducible findings with provable type-I error guarantees.

IV. AUTHOR CONTRIBUTIONS

M. S. developed the methodology, designed the experiments and performed the data analysis. E. K. contributed to some aspects of the analysis (quality control, phenotype extraction and SNP annotations) and the development of the visualization tool. S. B. contributed to the data acquisition. E. C. and C. S. supervised this project. All authors contributed to writing this manuscript.

V. ACKNOWLEDGEMENTS

We conducted this research using the UK Biobank Resource (application number 27837). M. S., E. C. and C. S. are supported by NSF grant DMS 1712800. E. C. and C. S. are also supported by a Math+X grant (Simons Foundation). E. K. is supported by the Hertz Foundation. S. B. is supported by a Ric Weiland fellowship. Finally, we thank Sahitya Mantravadi, Ananthakrishnan Ganesan, Robert Tibshirani and Trevor Hastie.

[1] Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

[2] Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* (2019). URL `https://doi.org/10.1038/s41576-019-0127-1`.

[3] Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).

[4] Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

[5] Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

[6] Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).

[7] Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284 (2015).

[8] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

[9] Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).

[10] Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).

[11] Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

[12] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

[13] Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

[14] Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).

[15] Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).

[16] Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

[17] Wang, G., Sarkar, A. K., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv* (2018).

[18] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.* **57**, 289–300 (1995).

[19] Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

[20] Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).

[21] Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

[22] O'Connell, J. *et al.* Haplotype estimation for biobank scale datasets. *Nat. Genet.* **48**, 817–820 (2016).

[23] Candès, E. J., Fan, Y., Janson, L. & Lv, J. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *J. R. Stat. Soc. B.* **80**, 551–577 (2018).

[24] Sesia, M., Sabatti, C. & Candès, E. J. Gene hunting with hidden markov model knockoffs. *Biometrika* **106**, 1–18 (2019).

[25] Bottolo, L. & Richardson, S. Discussion of gene hunting with hidden markov model knockoffs. *Biometrika* **106**, 19–22 (2019).

[26] Jewell, S. W. & Witten, D. M. Discussion of gene hunting with hidden markov model knockoffs. *Biometrika* **106**, 23–26 (2019).

[27] Rosenblatt, J. D., Ritov, Y. & Goeman, J. J. Discussion of gene hunting with hidden markov model knockoffs. *Biometrika* **106**, 29–33 (2019).

[28] Marchini, J. L. Discussion of gene hunting with hidden markov model knockoffs. *Biometrika* **106**, 27–28 (2019).

[29] Sesia, M., Sabatti, C. & Candès, E. J. Rejoinder: Gene hunting with hidden markov model knockoffs. *Biometrika* **106**, 35–45 (2019).

[30] Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

[31] Sabatti, C. *Multivariate Linear Models for GWAS*, 188–207 (Cambridge University Press, 2013).

[32] Davidson, I. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. *European Conference on Principles of Data Mining and Knowledge Discovery* 59–70 (2005).

[33] Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A. & Ron, M. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150**, 1699–1706 (1998).

[34] Sabatti, C., Service, S. & Freimer, N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* **164**, 829–833 (2003).

[35] Brzyski, D. *et al.* Controlling the rate of GWAS false discoveries. *Genetics* **205**, 61–75 (2017).

[36] Barber, R. F. & Candès, E. J. Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085 (2015).

[37] Dai, R. & Barber, R. F. The knockoff filter for FDR control in group-sparse and multitask regression. *J. Mach. Learn. Res.* **48**, 1851–1859 (2016).

[38] Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R, packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).

[39] Katsevich, E. & Sabatti, C. Multilayer knockoff filter: controlled variable selection at multiple resolutions. *Ann. Appl. Stat.* **13**, 1–33 (2019).

[40] Klasen, J. R. *et al.* A multi-marker association method for genome-wide association studies without the need for population structure correction. *Nat. Commun.* **7** (2016).

[41] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

[42] Katsevich, E., Sabatti, C. & Bogomolov, M. Controlling FDR while highlighting distinct discoveries. *arXiv* (2018).

[43] Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).

[44] Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

[45] McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495 (2010).

[46] Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

[47] Falush, D., Stephens, M. & Prichard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–87 (2003).

[48] Tang, H., Coram, M., Wang, P., Zhu, X. & Risch, N. Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* **79**, 1–12 (2006).

[49] Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Cambridge University Press, 2010).

[50] Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS genetics* **4**, e1000130 (2008).

[51] Guan, Y., Stephens, M. *et al.* Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* **5**, 1780–1815 (2011).

[52] Buzdugan, L. *et al.* Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics* **32**, 1990–2000 (2016).

ONLINE METHODS

A. Formally defining the objective of *KnockoffZoom*

We observe a phenotype $Y \in \mathbb{R}$ and genotypes $X = (X_1, \ldots, X_p) \in \{0,1,2\}^p$ for each individual. We assume that the pairs $(X^{(i)}, Y^{(i)})_{i=1}^n$ corresponding to n subjects are independently sampled from some distribution P_{XY}. The goal is to infer how $P_{Y|X}$ depends on X, testing the conditional hypotheses defined below, without assuming anything else about this likelihood, or restricting the sizes of n and p. Later, we will describe how we can achieve this by leveraging prior knowledge of the genotype distribution P_X. Now, we formally define the hypotheses.

Let $\mathcal{G} = (G_1, \ldots, G_L)$ be a partition of $\{1, \ldots, p\}$ into L blocks, for some $L \leq p$. For any $g \leq L$, we say that the g-th group of variables $X_{G_g} = \{X_j \mid j \in G_g\}$ is null if Y is independent of X_{G_g} given X_{-G_g} (X_{-G_g} contains all variables except those in G_g). We denote by $\mathcal{H}_0 \subseteq \{1, \ldots, L\}$ the subset of null hypotheses that are true. Conversely, groups containing causal variants do not belong to $\mathcal{H}_0$. For example, if $P_{Y|X}$ is a linear model, $\mathcal{H}_0$ collects the groups in $\mathcal{G}$ whose true coefficients are all zero (if there are no perfectly correlated variables in different groups[39]). In general, we want to select a subset $\hat{S} \subseteq \{1, \ldots, L\}$ as large as possible and such that $\text{FDR} = \mathbb{E}\left[|\hat{S} \cap \mathcal{H}_0| / \max(1, |\hat{S}|)\right] \leq q$, for a fixed partition $\mathcal{G}$ of the variants. These conditional hypotheses generalize those defined earlier in the statistical literature,[23,24] which only considered the variables one-by-one. Our hypotheses are better suited for the analysis of GWAS data because they allow us to deal with LD without pruning the variants (Supplement, Section S1 A). As a comparison, the null statement of the typical marginal hypothesis in a GWAS is that Y is marginally independent of X_j, for a given SNP j.

B. The knockoffs methodology

Knockoffs[23] solve the problem defined above if P_X is known and tractable, as explained below. The idea is to augment the data with synthetic variables, one for each genetic variant. We know that the knockoffs are null because we create them without looking at Y. Moreover we construct them so that they behave similarly to the SNPs in null groups and can serve as negative controls. The original work considered explicitly only the case of a trivial partition $\mathcal{G}$ into p singletons,[23] but we extend it for our problem by leveraging some previous work along this direction.[37,39] Formally, we say that $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ is a group-knockoff of X for a partition $\mathcal{G}$ of $\{1, \ldots, p\}$ if two conditions are satisfied: (1) Y is independent of $\tilde{X}$ given X; (2) the joint distribution of $(X, \tilde{X})$ is unchanged when $\{X_j : j \in G\}$ is swapped with $\{\tilde{X}_j : j \in G\}$, for any group $G \in \mathcal{G}$. The second condition is generally difficult to satisfy (unless $\tilde{X} = X$, which yields no power), depending on the form of P_X.[23] In the Supplement (Section S2), we develop algorithms to generate powerful group-knockoffs when P_X is an HMM. Here, we take $\tilde{X}$ as given and discuss how to test the conditional hypotheses. For the g-th group in $\mathcal{G}$, we compute feature importance measures T_g and $\tilde{T}_g$ for $\{X_j : j \in G_g\}$ and $\{\tilde{X}_j : j \in G_g\}$, respectively. Concretely, we fit a sparse linear (or logistic) regression model[38] for Y given $[X, \tilde{X}] \in \mathbb{R}^{n \times 2p}$, standardizing X and $\tilde{X}$; then we define $T_g = \sum_{j \in G_g} |\hat{\beta}_j(\lambda_{\text{CV}})|$, $\tilde{T}_g = \sum_{j \in G_g} |\hat{\beta}_{j+p}(\lambda_{\text{CV}})|$. Above, $\hat{\beta}_j(\lambda_{\text{CV}})$ and $\hat{\beta}_{j+p}(\lambda_{\text{CV}})$ indicate the estimated coefficients for X_j and $\tilde{X}_j$, respectively, with regularization parameter λ_{CV} tuned by cross-validation. These statistics are designed to detect sparse signals in a generalized linear model—a popular approximation of the distribution of Y in a GWAS.[31] Our power may be affected if this model is misspecified but our inferences remain valid. A variety of other tools could be used to compute more flexible or powerful statistics, perhaps incorporating prior knowledge.[23] Finally, we combine the importance measures into test statistics W_g through an anti-symmetric function, e.g., $W_g = T_g - \tilde{T}_g$, and report groups of SNPs with sufficiently large statistics.[23] The appropriate threshold for FDR control is calculated by the knockoff filter.[36] Further details about the test statistics are in the Supplement (Section S3). As currently implemented, our procedure has no power at the nominal FDR level q if there are fewer than $1/q$ findings to be made. Usually, this is not a problem for the analysis of complex traits, where many loci are significant. However, this may explain why, at the FDR level $q = 0.1$, we report none of the 5 discoveries obtained by BOLT-LMM for *glaucoma* in Table I. Alternatively, our method may detect these by slightly relaxing the knockoff filter,[36] at the cost of losing the provable FDR guarantee.

C. Quality control and data pre-processing for the UK Biobank

We consider 430,287 genotyped and phased subjects with British ancestry. According to the UK Biobank, 147,718 of these have at least one close relative in the dataset; we keep one from each of the 60,169 familial groups, chosen to minimize the missing phenotypes. This yields 350,119 unrelated subjects. We only analyze biallelic SNPs with minor allele frequency above 0.1% and in Hardy-Weinberg equilibrium (10^{-6}), among our 350,119 individuals. The final SNPs count is 591,513. A few subjects withdrew consent and we removed their observations from the analysis.

D. Including additional covariates

We control for the sex, age and squared age of the subjects to increase power (squared age is not used for *height*, as in earlier work[8]). We leverage these covariates by including them in the predictive model for the *KnockoffZoom* test statistics. We fit a sparse regression model on the augmented matrix of explanatory variables $[Z, X, \tilde{X}] \in \mathbb{R}^{n\times(m+2p)}$, where $Z, X, \tilde{X}$ contain the m covariates, the genotypes and their knockoff copies, respectively. The coefficients for Z are not regularized and we ignore them in the final computation of the test statistics.

Multi-resolution localization of causal variants across the genome

Supplementary Material

Matteo Sesia, Eugene Katsevich, Stephen Bates, Emmanuel Candès, Chiara Sabatti

Stanford University, Department of Statistics, Stanford, CA 94305, USA

CONTENTS

S1. METHODS

A. Knockoffs for composite conditional hypotheses

A typical GWAS involves so many variants in LD that the conditional importance of a single SNP given all others may be very difficult to resolve. We simplify this problem by clustering the loci into LD blocks and testing group-wise conditional hypotheses (Online Methods A). In particular, we test whether all variants in any given block are independent of the trait conditional on the rest of the genome.

Consider the following explanatory example. A trait Y is described by a linear model (for simplicity) with 4 explanatory variables: $Y = \sum_{j=1}^{4} X_j \beta_j + \epsilon$ and Gaussian noise ϵ. Given independent observations of X and Y, drawn with $\beta_1 = 1$ and $\beta_j = 0$ for all $j \neq 1$, we want to discover which variables influence Y (i.e., X_1). To make this example interesting, we imagine an extreme form of LD: $X_1 = X_2$ and $X_3 = X_4$, while X_1 and X_3 are independent. This makes it impossible to retrospectively understand whether it is X_1 or X_2 that affects Y. However, we can still hope to conclude that either X_1 or X_2 are important. In fact, introductory statistics classes teach us how to do this with an F-test, as opposed to a t-test, which would have no power in this case.

To solve the above problem within our framework (Online Methods B), we begin by generating powerful knockoffs specifically designed to test group-wise conditional hypotheses. The exact structure of the LD blocks must be taken into account when we create knockoffs. For example, if we naïvely generate knockoffs $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_4)$ that are pairwise exchangeable with X one-by-one, we must set $\tilde{X}_1 = X_2 = X_1$ to preserve the equality in distribution between $(\tilde{X}_1, X_2, X_3, X_4)$ and $(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_4)$ when $\tilde{X}_1$ is swapped with X_1. Furthermore, by the same argument we conclude that we need $\tilde{X} = X$. However, these knockoffs are powerless as negative controls (although they are exchangeable). Fortunately, we can generate powerful group-knockoffs (Online Methods B) under milder exchangeability constraints that can still be used to test grouped hypotheses. In the above toy example, we consider the partition $\mathcal{G} = (\{1, 2\}, \{3, 4\})$, and require that $\tilde{X}_1 = \tilde{X}_2$ and $\tilde{X}_3 = \tilde{X}_4$, while allowing $\tilde{X} \neq X$.

Definition 1 (Group-knockoffs). *Consider random variables $Z = (Z_1, \ldots, Z_p)$, and a partition $\mathcal{G} = (G_1, \ldots, G_L)$ of $\{1, \ldots, p\}$. Then $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_p)$ is said to be a* group-knockoff *for Z with respect to $\mathcal{G}$ if for each group $G \in \mathcal{G}$, we have:*

$$(Z, \tilde{Z})_{swap(G;\mathcal{G})} \overset{d}{=} (Z, \tilde{Z}). \tag{S1}$$

Above, $(Z, \tilde{Z})_{swap(G;\mathcal{G})}$ means that the jth coordinate is swapped with the $(j+p)$th coordinate $\forall j \in G$.

B. Coordinating discoveries across resolutions

In this section, we show how to combine the tests statistics computed by our method at different resolutions in order to coordinate the discoveries, so that no "floating" blocks are reported, while rigorously controlling the FDR.

Let $r = 1, \dots, R$ index the resolutions we consider (Online Methods A), ordered from the lowest to the highest. At each resolution r, the p variants are partitioned into L^r groups: $\mathcal{G}^r = (G^r_1, \dots, G^r_{L^r})$. By construction, the partitions are nested, so that for each resolution $r > 1$ and each group $g \in \{1, \dots, L^r\}$ there is a unique parent group $\text{Pa}(g,r) \in \{1, \dots, L^{r-1}\}$ such that $G^r_g \subseteq G^{r-1}_{\text{Pa}(g,r)}$. Suppose that we have computed the test statistics $W^r = \{W^r_g\}_{g=1,\dots,L^r}$ at each resolution (Online Methods B). To avoid "floating" discoveries, we can only select groups whose parent has been discovered at the resolution below. We can enforce this consistency property while preserving an FDR guarantee at each resolution by slightly modifying the final filtering step that computes the significance thresholds (Online Methods B). Instead of applying the knockoff filter separately at each resolution, we proceed with Algorithm 1 sequentially, from lower to higher resolutions.

Algorithm 1 Multi-resolution knockoff filter

```
Input: Partitions $\mathcal{G}^r$, for each resolution level $r \in \{1, \dots, R\}$;
       Test statistics $W^r = (W^r_1, \dots, W^r_{L^r})$, for each resolution level $r \in \{1, \dots, R\}$;
       FDR target level $q$.
Compute $t^{*1}$ applying the usual knockoff filter, at nominal level $q/1.93$, to $W^1$:
    $t^{*1} = \min \left\{ t^1 \geq 0 : \frac{1 + |\{g : W^1_g \leq -t^1\}|}{|\{g : W^1_g \geq t^1\}|} \leq \frac{q}{1.93} \right\}.$
Select discoveries at the lowest resolution: $\hat{S}^1 = \{g : W^1_g \geq t^{*1}\}$.
for $r = 2$ to $r = R$ do
    Compute $t^{*r}$ as follows:
        $t^{*r} = \min \left\{ t^r \geq 0 : \frac{1 + |\{g : W^r_g \leq -t^r\}|}{|\{g : W^r_g \geq t^r,\ \text{Pa}(g,r) \in \hat{S}^{r-1}\}|} \leq \frac{q}{1.93} \right\}.$
    Select discoveries at this resolution: $\hat{S}^r = \{g : W^r_g \geq t^{*r},\ \text{Pa}(g,r) \in \hat{S}^{r-1}\}$.
```

We prove in Proposition 1 that the FDR is controlled at each resolution by Algorithm 1, which is closely inspired by previous work.[1] The correction of the FDR level by the factor 1.93 is required by the proof for technical reasons, although it has been observed empirically in a similar context[1] that this may be practically unnecessary. Therefore, to avoid an unjustified power loss while retaining provable guarantees, we prefer not to apply Algorithm 1 in this work. However, should "floating" discoveries become particularly undesirable, we suspect that it may be safe to apply Algorithm 1 even without the 1.93 factor.

Proposition 1. *Denote by* FDR^r *the FDR at resolution* r *for the discoveries obtained with Algorithm 1. If the data sampling assumptions (Online Methods A) are valid and the knockoff exchangeability holds (Online Methods B), then* $\mathrm{FDR}^r \leq 1.93 \cdot q$, $\forall r \in \{1, \ldots, R\}$.

Proof. It was shown in earlier work[1] (proof of Theorem 1 therein) that

$$\mathbb{E}\left[\sup_{t^r \geq 0} \frac{|\{g : W_g^r \geq t^r\} \cap \mathcal{H}_0^r|}{1 + |\{g : W_g^r \leq -t^r\}|}\right] \leq 1.93,$$

where $\mathcal{H}_0^r \subseteq \{1, \ldots, L^r\}$ is the set of null groups at resolution r. Therefore,

$$\begin{aligned}
\mathrm{FDR}^r = \mathbb{E}\left[\frac{|\hat{S}^r \cap \mathcal{H}_0^r|}{|\hat{S}^r|}\right] &\leq \mathbb{E}\left[\frac{|\{g : W_g^r \geq t^{*r}\} \cap \mathcal{H}_0^r|}{|\hat{S}^r|}\right] \\
&= \mathbb{E}\left[\frac{|\{g : W_g^r \geq t^{*r}\} \cap \mathcal{H}_0^r|}{1 + |\{g : W_g^r \leq -t^{*r}\}|} \cdot \frac{1 + |\{g : W_g^r \leq -t^{*r}\}|}{|\hat{S}^r|}\right] \\
&\leq \frac{q}{1.93} \cdot \mathbb{E}\left[\frac{|\{g : W_g^r \geq t^{*r}\} \cap \mathcal{H}_0^r|}{1 + |\{g : W_g^r \leq -t^{*r}\}|}\right] \\
&\leq \frac{q}{1.93} \cdot \mathbb{E}\left[\sup_{t^r \geq 0} \frac{|\{g : W_g^r \geq t^r\} \cap \mathcal{H}_0^r|}{1 + |\{g : W_g^r \leq -t^r\}|}\right] \\
&\leq q.
\end{aligned}$$

□

S2. ALGORITHMS FOR KNOCKOFF GENERATION

The construction of knockoffs for groups of genetic variants is the heart of the methodology in this paper and requires a significant extension of the existing algorithms.[2] Our contribution has two main components: first, we construct group-knockoffs for Markov chains and HMMs; second, we specialize these ideas to obtain fast algorithms for the models of interest in a GWAS. The details of these contributions are in this section, while the associated technical proofs are in Section S6.

A. General algorithms for group-knockoffs

1. Markov chains

We begin with group-knockoffs for Markov chains.[2] We say that $Z = (Z_1, \ldots, Z_p)$, with each variable taking values in $\{1, \ldots, K\}$ for some $K \in \mathbb{N}$, is a discrete Markov chain if its joint probability mass function can be written as:

$$\mathbb{P}[Z_1 = z_1, \dots, Z_p = z_p] = Q_1(z_1) \prod_{j=2}^{p} Q_j(z_j \mid z_{j-1}). \tag{S2}$$

Above, $Q_1(z_1) = \mathbb{P}[Z_1 = z_1]$ denotes the initial distribution of the chain, while the transition matrices between consecutive variables are: $Q_j(z_j \mid z_{j-1}) = \mathbb{P}[Z_j = z_j \mid Z_{j-1} = z_{j-1}]$.

Since Markov chains have a well-defined sequential structure, the special class of contiguous partitions is of particular interest for generating group-knockoffs.

Definition 2 (Contiguous partition). *For any fixed positive integers $L \leq p$, we call a collection of L sets $\mathcal{G} = (G_1, \dots, G_L)$ a* contiguous partition *of $\{1, \dots, p\}$ if $\mathcal{G}$ is a partition of $\{1, \dots, p\}$ and for any distinct $G, G' \in \mathcal{G}$, either $j < l$ for all $j \in G$ and $l \in G'$ or $j > l$ for all $j \in G$ and $l \in G'$.*

For example, the partition $\mathcal{G} = (G_1, \dots, G_4)$ of $\{1, \dots, 10\}$ shown on the left-hand-side of the example below is contiguous, while $\mathcal{G}' = (G'_1, \dots, G'_3)$, on the right-hand-side, is not.

$$\underbrace{1,2,3}_{G_1}, \underbrace{4,5}_{G_2}, \underbrace{6,7}_{G_3}, \underbrace{8,9,10}_{G_4}, \qquad\qquad \underbrace{1,2,3}_{G'_1}, \underbrace{4,5}_{G'_2}, \underbrace{6,7}_{G'_3}, \underbrace{8,9,10}_{G'_2}.$$

We consider only contiguous partitions when we construct knockoff copies of a Markov chain; if a given partition is not contiguous, we first refine it by splitting all non-contiguous groups.

To simplify the notation in the upcoming result, for any $g \in \{1, \dots, L\}$ and associated group $G_g \in \mathcal{G}$, we indicate the variables in G_g as: $Z^g = (Z_1^g, \dots, Z_{m_g}^g) = (Z_j)_{j \in G_g}$. Similarly, we denote the sequence of transition matrices corresponding to the m_g variables contained in the g-th group by: $Q^g = (Q_1^g, \dots, Q_{m_g}^g) = (Q_j)_{j \in G_g}$. We set $Q_1^1(k \mid l) = Q_1(k)$ and $Q_1^{L+1}(k \mid l) = 1$, $\forall k, l \leq K$.

Proposition 2. *Let $\mathcal{G} = (G_1, \dots, G_L)$ be a contiguous partition of $\{1, \dots, p\}$, such that the g-th group G_g has m_g elements. Suppose that Z is distributed as the Markov chain in* (S2)*, with known parameters Q. Then, a group-knockoff copy $\tilde{Z}$, with respect to $\mathcal{G}$, can be obtained by sequentially sampling, for $g = 1, \dots, L$, the g-th group-knockoff copy $\tilde{Z}^g = (\tilde{Z}_1^g, \dots, \tilde{Z}_{m_g}^g)$ from:*

$$\begin{aligned} p(Z^g \mid Z^{-g}, \tilde{Z}^{1:(g-1)}) = \frac{1}{\mathcal{N}_g(Z_1^{g+1})} &\times \frac{Q_1^g(Z_1^g \mid \tilde{Z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g \neq 1]} \, Q_1^g(Z_1^g \mid Z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(Z_1^g)} \\ &\times \left[\prod_{j=2}^{m_g} Q_j^g(Z_j^g \mid Z_{j-1}^g) \right] \times Q_1^{g+1}(Z_1^{g+1} \mid Z_{m_g}^g). \end{aligned} \tag{S3}$$

The functions $\mathcal{N}$ are defined recursively as:

$$\mathcal{N}_g(k) = \sum_{z_1^g, \dots, z_{m_g}^g} \frac{Q_1^g(Z_1^g \mid \tilde{Z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g \neq 1]} \, Q_1^g(Z_1^g \mid Z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(Z_1^g)} \times \left[\prod_{j=2}^{m_g} Q_j^g(z_j^g \mid z_{j-1}^g) \right] \times Q_1^{g+1}(k \mid z_{m_g}^g), \tag{S4}$$

with the convention that $\mathcal{N}_0(z) = 1$ for all z. Therefore, Algorithm 2 is an exact procedure for sampling group-knockoff copies of a Markov chain.

Algorithm 2 Group-knockoffs for a Markov chain

```
for $g = 1$ to $g = L$ do
  for $k = 1$ to $k = K$ do
    Compute $\mathcal{N}_g(k)$ according to (S4).
    Sample $\tilde{Z}^g$ according to (S3).
```

Algorithm 2 reduces to the known result for ungrouped knockoffs[2] if $L = p$. In the general case, we can sample from the distribution in (S3) as follows. From (S3), we can write:

$$p(\tilde{Z}^g \mid Z^{-g}, \tilde{Z}^{1:(g-1)}) = Q_1^{\star g}(\tilde{Z}_1^g) \prod_{j=2}^{m_g} Q_j^{\star g}(\tilde{Z}_j^g \mid \tilde{Z}_{j-1}^g),$$

for some initial distribution $Q_1^{\star g}$ and suitable transition matrices $Q_j^{\star g}$. Therefore, $\tilde{Z}^g$ is conditionally a Markov chain. Assuming for simplicity that $g > 1$, we see from (S3) that $Q_1^{\star g}$ is given by:

$$\begin{aligned} Q_1^{\star g}(\tilde{z}_1^g) = & \frac{1}{\mathcal{N}_g(Z_1^{g+1})} \times \frac{Q_1^g(\tilde{z}_1^g \mid \tilde{Z}_{m_{g-1}}^{g-1})\, Q_1^g(\tilde{z}_1^g \mid Z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(\tilde{z}_1^g)} \\ & \times \left(\sum_{\tilde{z}_2^g, \ldots, \tilde{z}_{m_g}^g} \left[\prod_{j=2}^{m_g} Q_j^g(\tilde{z}_j^g \mid \tilde{z}_{j-1}^g) \right] \times Q_1^{g+1}(Z_1^{g+1} \mid \tilde{z}_{m_g}^g) \right), \end{aligned}$$

while, for $j \in \{2, \ldots, m_g\}$,

$$Q_j^{\star g}(\tilde{z}_j^g \mid \tilde{z}_{j-1}^g) = Q_j^g(\tilde{z}_j^g \mid \tilde{z}_{j-1}^g) \times \left(\sum_{\tilde{z}_{j+1}^g, \ldots, \tilde{z}_{m_g}^g} \left[\prod_{j'=j+1}^{m_g} Q_{j'}^g(\tilde{z}_{j'}^g \mid \tilde{z}_{j'-1}^g) \right] \times Q_1^{g+1}(Z_1^{g+1} \mid \tilde{z}_{m_g}^g) \right).$$

It is easy to verify that the above quantities can be computed through $m_g - 1$ multiplications of $K \times K$ matrices. Similarly, the functions in (S4) can also be computed efficiently. Therefore, the cost of sampling the g-th group-knockoff is $\mathcal{O}(m_g K^3)$, if $m_g > 1$, and $\mathcal{O}(m_g K^2)$, otherwise. The worst-case complexity of Algorithm 2 is $\mathcal{O}(pK^3)$ because $\sum_{g=1}^{L} m_g = p$. Later, we will derive a more efficient implementation in the special case of genetic variables.

2. *Hidden Markov models*

We leverage the result in Proposition 2 to derive a construction of group-knockoffs for HMMs, similarly to previous work.[2] We say that $X = (X_1, \ldots, X_p)$, with each variable taking values in a finite state space $\mathcal{X}$, is distributed as an HMM with K hidden states if there exists a vector of latent random variables $Z = (Z_1, \dots, Z_p)$, with $Z_j \in \{1, \dots, K\}$, such that:

$$\begin{cases} Z \sim \text{MC}\,(Q) & \text{(latent discrete Markov chain)}, \\ X_j \mid Z \sim X_j \mid Z_j \overset{\text{ind.}}{\sim} f_j(X_j \mid Z_j) & \text{(emission distribution)}. \end{cases} \tag{S5}$$

Above, MC$\,(Q)$ indicates the law of a discrete Markov chain, as in (S2).

Proposition 3. *Suppose $X = (X_1, \dots, X_p)$ is distributed as the HMM in* (S5)*, with an associated latent Markov chain $Z = (Z_1, \dots, Z_p)$. Let $\mathcal{G} = (G_1, \dots, G_L)$ be a contiguous partition of $\{1, \dots, p\}$. Then, Algorithm 3 generates $(\tilde{X}, \tilde{Z})$ such that:*

$$\left((X, \tilde{X})_{swap(G;\mathcal{G})}, (Z, \tilde{Z})_{swap(G;\mathcal{G})}\right) \overset{d}{=} \left((X, \tilde{X}), (Z, \tilde{Z})\right), \qquad \forall G \in \mathcal{G}. \tag{S6}$$

In particular, this implies that $\tilde{X}$ is a group-knockoff copy of X.

```
Algorithm 3 Group-knockoffs for an HMM
  (1) Sample $Z = (Z_1, \dots, Z_p)$ from $\mathbb{P}[Z \mid X = x]$ using forward-backward sampling.[2]
  (2) Sample a group-knockoff copy $\tilde{Z}$ of $Z$, with respect to $\mathcal{G}$, using Algorithm 2.
  (3) Sample $\tilde{X}$ from $\mathbb{P}[X \mid Z = \tilde{z}]$. This is trivial by the conditional independence in (S5).
```

The computational complexity of the first step of Algorithm 3 is known to be $\mathcal{O}(pK^2)$,[2] while the worst-case cost of the second step is $\mathcal{O}(pK^3)$. The complexity of the third step is $\mathcal{O}(p|\mathcal{X}|)$. Therefore, the worst-case total complexity of Algorithm 3 is $\mathcal{O}(p(K^3 + |\mathcal{X}|))$. Later, we will reduce this in the special case of the Li-and-Stephens HMM by simplifying the first two steps analytically.

B. The Li-and-Stephens model

1. *Model parameterization*

We specialize Algorithm 3 in the case of the Li-and-Stephens model.[3] This HMM describes the distribution of genotypes as a patchwork of latent ancestral motifs. A quick preview of the results: in this section we introduce the model; in Section S2 C we optimize the second step of Algorithm 3 to have complexity $\mathcal{O}(pK)$ for phased haplotypes, or $\mathcal{O}(pK^2)$ for unphased genotypes; and in Section S2 D we optimize the first step of Algorithm 3 to have complexity $\mathcal{O}(pK)$ for phased haplotypes, or $\mathcal{O}(pK^2)$ for unphased genotypes. By combining these results, we decrease the complexity of Algorithm 3 to $\mathcal{O}(pK)$ for phased haplotypes, or $\mathcal{O}(pK^2)$ for unphased genotypes. This is an important contribution because it makes *KnockoffZoom* applicable to large datasets.

The Li-and-Stephens model for phased haplotype sequences describes $X = (X_1, \dots, X_p)$, with $X_j \in \{0, 1\}$, as an imperfect mosaic of K ancestral motifs, $D_i = (D_{i,1}, \dots, D_{i,p}\}$, for $i \in \{1, \dots, K\}$ and $D_{i,j} \in \{0, 1\}$. This can be formalized as an HMM with K hidden states, in the form of (S5).[4] The transitions of the latent Markov chain are simple:

$$\mathbb{P}\left[Z_j = k \mid Z_{j-1} = l\right] = Q_j(k \mid l) = a_{j,k} + b_j \delta_{k,l}. \tag{S7}$$

Above, $\delta_{k,l}$ indicates the Kronecker delta: $\delta_{k,l}$ is equal to 1 if $k = l$, and 0 otherwise. Conditional on Z, each X_j is drawn independently from:

$$\mathbb{P}\left[X_j = 1 \mid Z\right] = \mathbb{P}\left[X_j = 1 \mid Z_j\right] = \theta_{j,Z_j}.$$

The parameters $\theta = (\theta_{j,k})_{k \in [K], j \in [p]}$ describe the haplotype motifs in D and the mutation rates. We can write a and b consistent with the notation of fastPHASE[4] as:

$$a_{j,u} = \begin{cases} \alpha_{1,u}, & \text{if } j = 1, \\ \left(1 - e^{-r_j}\right) \alpha_{j,u}, & \text{if } j > 1, \end{cases} \qquad b_j = \begin{cases} 0, & \text{if } j = 1, \\ e^{-r_j}, & \text{if } j > 1. \end{cases}$$

The parameters $\alpha = (\alpha_{j,k})_{k \in [K], j \in [p]}$ describe the prevalence of each motif in the population. The likelihood of a transition in the Markov chain depends on the values of $r = (r_1, \dots, r_p)$, which capture the genetic recombination rates along the genome. This phenomenological model of LD has inspired several successful applications for phasing and imputation.[5]

An unphased genotype sequence $X = (X_1, \dots, X_p)$, with $X_j \in \{0, 1, 2\}$, can be described as the element-wise sum of two independent and identically distributed haplotype sequences, H^a and H^b, that follow the model defined above. Consequently, X is also distributed as an HMM in the form of (S5) with $K_{\text{eff}} = K(K + 1)/2$ hidden states, where K is the number of haplotype motifs. This quadratic dependence on K follows from the fact that each latent Markov state of X corresponds to an unordered pair of states, $\{Z_j^a, Z_j^b\}$, corresponding to the unobserved haplotypes. The transition probabilities for the effective Markov chain have the following structure:

$$\begin{aligned} \mathbb{P}\left[Z_j = \{k^a, k^b\} \mid Z_{j-1} = \{l^a, l^b\}\right] &= \bar{Q}_j(\{k^a, k^b\} \mid \{l^a, l^b\}) \\ &= \begin{cases} Q_j(k^a \mid l^a)\, Q_j(k^b \mid l^b), & \text{if } k^a = k^b, \\ Q_j(k^a \mid l^a)\, Q_j(k^b \mid l^b) + Q_j(k^b \mid l^a)\, Q_j(k^a \mid l^b), & \text{otherwise.} \end{cases} \end{aligned} \tag{S8}$$

Above, $Q_j(k \mid l)$ is given by (S7) while the conditional emission distributions are:

$$\mathbb{P}\left[X_j = x \mid Z = z\right] = \mathbb{P}\left[X_j = x \mid Z_j = \{k^a, k^b\}\right] = \begin{cases} (1-\theta_{j,k^a})(1-\theta_{j,k^b}), & \text{if } x = 0, \\ \theta_{j,k^a}(1-\theta_{j,k^b}) + (1-\theta_{j,k^a})\theta_{j,k^b}, & \text{if } x = 1, \\ \theta_{j,k^a}\theta_{j,k^b}, & \text{if } x = 2. \end{cases}$$

Generating group-knockoffs for genotypes using the algorithms in Section S2 would cost $\mathcal{O}(pK^6)$, while knockoffs for haplotypes would cost $\mathcal{O}(pK^3)$. This is prohibitive for large datasets, since the operation must be repeated separately for each subject. Before proceeding to simplify the algorithm analytically, we state the following useful lemma.

Lemma 1. *The Markov chain transition matrices for the unphased genotypes in the Li-and-Stephens HMM can be written explicitly as:*

$$\begin{aligned}\bar{Q}_j(\{k^a, k^b\} \mid \{l^a, l^b\}) = a_{j,k^a} a_{j,k^b}(2 - \delta_{k^a,k^b}) + (b_j)^2 \delta_{\{l^a,l^b\},\{k^a,k^b\}} \\ + b_j \frac{a_{j,k^a}\left(\delta_{k^b,l^a} + \delta_{k^b,l^b}\right) + a_{j,k^b}\left(\delta_{k^a,l^a} + \delta_{k^a,l^b}\right)}{1 + \delta_{k^a,k^b}}.\end{aligned}$$

2. *Utilizing phased haplotypes*

We leverage the phased haplotypes in the UK Biobank data[6] to accelerate the generation of the group-knockoffs for genotypes. Denote the haplotypes of one subject as $H^a, H^b \in \{0,1\}^p$, so that the genotypes are $X = H^a + H^b \in \{0,1,2\}^p$. Only X is measures in a GWAS, whereas H^a and H^b are probabilistic reconstructions based on an HMM[5] similar to that in Section S2 B 1. Holding this thought, note that the following algorithm generates exact group-knockoffs for the genotypes: first, sample $\{H^a, H^b\}$ from their posterior distribution given X; next, create group-knockoffs $\tilde{H}^a$ for H^a and $\tilde{H}^b$ for H^b, independently; and lastly, set $\tilde{X} = \tilde{H}^a + \tilde{H}^b$. The proof is equivalent to that of Proposition 3. The first step above corresponds to phasing (although sometimes phasing is carried out by reconstructing the most likely haplotypes H^a and H^b, as opposed to posterior sampling). Given the reconstructed haplotypes, the second stage of the above algorithm only involves an HMM with K latent states, instead of $\mathcal{O}(K^2)$. The third stage of the algorithm is trivial.

C. Efficient sampling of Markov chain knockoffs

1. *Phased haplotypes*

We begin to specialize Algorithm 2 for the HMM of phased haplotypes in Section S2 B, starting from its second step. Recall that the $\mathcal{N}$ functions in (S4) are defined recursively as:

$$\mathcal{N}_g(k) = \sum_{z_1^g,\dots,z_{m_g}^g} \frac{Q_1^g(z_1^g \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}\, Q_1^g(z_1^g \mid z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(z_1^g)} \times \left[\prod_{j=2}^{m_g} Q_j^g(z_j^g \mid z_{j-1}^g)\right] \times Q_1^{g+1}(k \mid z_{m_g}^g).$$

To simplify the notations, define, for each group g, $v^g \in \mathbb{R}^{m_g \times K}$ and $u^g \in \mathbb{R}^{m_g}$ as follows:

- The last row of v^g and the last element of u^g are:

$$v_{m_g,k}^g = a_{1,k}^{g+1}, \quad \forall k \in \{1,\dots,K\}, \qquad\qquad u_{m_g}^g = b_1^{g+1}.$$

- For $j \in \{1,\dots,m_g-1\}$, the j-th row of v^g and the j-th element of u^g are defined recursively:

$$v_{j,k}^g = v_{j+1,k}^g + u_{j+1}^g a_{j+1,k}^g, \quad \forall k \in \{1,\dots,K\}, \qquad\qquad u_j^g = u_{j+1}^g b_{j+1}^g.$$

Proposition 4. *For the special case of the Li-and-Stephens model of unphased genotypes, the $\mathcal{N}$ function in Algorithm 2 for the g-th group* (S4) *can be computed recursively as:*

$$\mathcal{N}_g(k) = u_1^g \frac{Q_1^g(k \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}\, Q_1^g(k \mid z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(k)} + v_{1,k}^g \sum_{l=1}^{K} \frac{Q_1^g(l \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}\, Q_1^g(l \mid z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(l)}. \tag{S9}$$

Above, it is understood that $\mathcal{N}_0(k) = 1$, $\forall k$.

Each $v_j^g \in \mathbb{R}^K$ can be computed in $\mathcal{O}(K)$ time, while the additional cost for u_j^g is $\mathcal{O}(1)$. Therefore, we compute v_1^g and u_1^g in $\mathcal{O}(m_g K)$ time and evaluate $\mathcal{N}_g(k)$ in $\mathcal{O}(m_g K)$ time (Proposition 4).

Given $\mathcal{N}_g$, we must sample the vector $\tilde{Z}^g = (\tilde{Z}_1^g,\dots,\tilde{Z}_{m_g}^g)$ from:

$$\mathbb{P}\left[\tilde{Z}^g = \tilde{z}^g \mid z^{-g}, \tilde{z}^{1:(g-1)}\right] \propto \frac{Q_1^g(\tilde{z}_1^g \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}\, Q_1^g(\tilde{z}_1^g \mid z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(\tilde{z}_1^g)} \times \left[\prod_{j=2}^{m_g} Q_j^g(\tilde{z}_j^g \mid \tilde{z}_{j-1}^g)\right] \times Q_1^{g+1}(z_1^{g+1} \mid \tilde{z}_{m_g}^g).$$

This is a multivariate distribution from which we can sample efficiently, as stated next.

Proposition 5. *In the special case of Algorithm 2 for the Li-and-Stephens model of phased haplotypes, the knockoff copy for the first element of the g-th group can be sampled from:*

$$\mathbb{P}\left[\tilde{Z}_1^g = k \mid z^{-g}, \tilde{z}^{1:(g-1)}\right] \propto \frac{Q_1^g(k \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}\, Q_1^g(k \mid z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(k)} \left(v^g_{1,z_1^{g+1}} + u_1^g \mathbb{1}\left[k = z_1^{g+1}\right]\right).$$

For $j \in \{2, \ldots, m_g\}$, the knockoff copy for the j-th element of the g-th group can be sampled from:

$$\mathbb{P}\left[\tilde{Z}_j^g = k \mid z^{-g}, \tilde{z}^{1:(g-1)}, \tilde{z}^g_{1:(j-1)}\right] \propto Q_j^g(k \mid \tilde{z}_{j-1}^g) \left(v^g_{j,z_1^{g+1}} + u_j^g \mathbb{1}\left[k = z_1^{g+1}\right]\right).$$

Each coordinate is sampled at an additional cost $\mathcal{O}(K)$, reusing v_j^g, u_j^g from the computation of the $\mathcal{N}$ functions. The cost for the sequence is $\mathcal{O}(\sum_{g=1}^G m_g K) = \mathcal{O}(pK)$, regardless of the grouping.

2. *Unphased genotypes*

We perform analogous calculations in the case of the HMM for unphased genotypes. Here the notation is more involved. The functions in (S4) are defined recursively as:

$$\mathcal{N}_g(\{k^a, k^b\}) = \sum_{\{l_1^a, l_1^b\}, \ldots, \{l_{m_g}^a, l_{m_g}^b\}} \frac{\bar{Q}_1^g(\{l_1^a, l_1^b\} \mid z_{m_{g-1}}^{g-1})\, \bar{Q}_1^g(\{l_1^a, l_1^b\} \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}}{\mathcal{N}_{g-1}(\{l_1^a, l_1^b\})}$$
$$\times \left[\prod_{j=2}^{m_g} \bar{Q}_j^g(\{l_j^a, l_j^b\} \mid \{l_{j-1}^a, l_{j-1}^b\})\right] \times \bar{Q}_1^{g+1}(\{k^a, k^b\} \mid \{l_{m_g}^a, l_{m_g}^b\}).$$

We start by defining some new variables, as in Section S2 C 1. For $k, k^a, k^b \in \{1, \ldots, K\}$, let:

$$u_{m_g}^g = (b_1^{g+1})^2, \qquad v^g_{m_g,\{k^a,k^b\}} = a^{g+1}_{1,k^a} a^{g+1}_{1,k^b}, \qquad w^g_{m_g,k} = b_1^{g+1} a^{g+1}_{1,k}.$$

For $j \in \{1, \ldots, m_g - 1\}$ and $k^a, k^b \in \{1, \ldots, K\}$, we define recursively:

$$\begin{aligned}
u_j^g &= u_{j+1}^g (b_{j+1}^g)^2, \\
v^g_{j,\{k^a,k^b\}} &= v^g_{j+1,\{k^a,k^b\}} + w^g_{j+1,k^a} a^g_{j+1,k^b} + w^g_{j+1,k^b} a^g_{j+1,k^a} + u^g_{j+1} a^g_{j+1,k^a} a^g_{j+1,k^b}, \\
w^g_{j,k} &= w^g_{j+1,k} b^g_{j+1} + u^g_{j+1} a^g_{j+1,k} b^g_{j+1}.
\end{aligned}$$

Moreover, we also define:

$$C^g(\{k^a, k^b\}) = \frac{\bar{Q}_1^g(\{k^a, k^b\} \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}\, \bar{Q}_1^g(\{k^a, k^b\} \mid z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(\{k^a, k^b\})},$$

and

$$D^g(k^a) = C^g(k^a, k^a) + \sum_{k^b=1}^{K} C^g(\{k^a, k^b\}).$$

The following result shows that the cost of computing the g-th $\mathcal{N}$ function is $\mathcal{O}(m_g K^2)$.

Proposition 6. *For the special case of the Li-and-Stephens model of unphased genotypes, the $\mathcal{N}$ function in Algorithm 2 for the g-th group is given by:*

$$\begin{aligned}\mathcal{N}_g(\{k^a, k^b\}) = \frac{w^g_{1,k^a} D^g(k^b) + w^g_{1,k^b} D^g(k^a)}{1+\delta_{k^a,k^b}} &+ u^g_1 C^g(\{k^a, k^b\}) \\ &+ (2-\delta_{k^a,k^b}) v^g_{1,\{k^a,k^b\}} \sum_{\{l^a,l^b\}} C^g(\{l^a, l^b\}).\end{aligned}$$

The sampling part of Algorithm 2 is similar to that in Section S2 C 1.

Proposition 7. *For the special case of the Li-and-Stephens model of unphased genotypes, the knockoff copy for the first element of the g-th group can be sampled in Algorithm 2 from:*

$$\mathbb{P}\left[\tilde{Z}^g_1 = \{k^a, k^b\} \mid z^{-g}, \tilde{z}^{1:(g-1)}\right] \propto \frac{\bar{Q}^g_1(\{k^a, k^b\} \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g\neq 1]}\, \bar{Q}^g_1(\{k^a, k^b\} \mid z^{g-1}_{m_{g-1}})}{\mathcal{N}_{g-1}(\{k^a, k^b\})} V^g_1(z^{g+1}_1 \mid \{k^a, k^b\}).$$

The knockoff for the j-th element of the g-th group, for $j \in \{2, \dots, m_g\}$, can be sampled from:

$$\mathbb{P}\left[\tilde{Z}^g_j = \{k^a, k^b\} \mid z^{-g}, \tilde{z}^{1:(g-1)}\right] \propto \bar{Q}^g_j(\{k^a, k^b\} \mid \tilde{z}^g_{j-1})\, V^g_j(z^{g+1}_1 \mid \{k^a, k^b\}).$$

Above, the variables V^g_j are defined as:

$$\begin{aligned}V^g_j(\{l^a, l^b\} \mid \{k^a, k^b\}) = v^g_{j,\{l^a,l^b\}}(2-\delta_{l^a,l^b}) &+ u^g_j \delta_{\{l^a,l^b\},\{k^a,k^b\}} \\ &+ \frac{w^g_{j,l^a}\left(\delta_{l^b,k^a} + \delta_{l^b,k^b}\right) + w^g_{j,l^b}\left(\delta_{l^a,k^a} + \delta_{l^a,k^b}\right)}{1+\delta_{l^a,l^b}}.\end{aligned}$$

Sampling group-knockoffs costs $\mathcal{O}(\sum_{g=1}^G m_g K^2) = \mathcal{O}(pK^2)$ per individual, with any grouping.

D. Efficient forward-backward sampling

1. *Forward-backward sampling for HMMs*

The first step in Algorithm 3 consists of sampling Z from the posterior distribution of the latent Markov chain in (S5) given X. In general, this requires a variation of Viterbi's algorithm,[2] as recalled in Algorithms 4 and 5. Here, K indicates the number of states in the Markov chain.

Algorithm 4 Forward-backward sampling (forward pass)

```
Initialize $F_0 = 1$,   $Q_1(k \mid l) = Q_1(k)$, for all $k, l$
for $j = 1$ to $p$ do
    for $k = 1$ to $K$ do
        Compute $F_j(k) = f_j(x_j \mid k) \sum_{l=1}^K Q_j(k \mid l)\, F_{j-1}(l)$.
```

Algorithm 5 Forward-backward sampling (backward pass)

```
Initialize $j = p$,   $Q_{p+1}(k \mid l) = 1$ for all $k, l$
for $j = p$ to 1 (backward) do
    Sample $Z_j$ from $\mathbb{P}[Z_j = k] = \frac{Q_{j+1}(Z_{j+1}|k)\, F_j(k)}{\sum_{l=1}^{K} Q_{j+1}(Z_{j+1}|l)\, F_j(l)}$.
```

2. *Phased haplotypes*

The forward probabilities in Algorithm 4 are defined recursively, for $j \in \{2, \dots, p\}$, as:

$$\begin{aligned} F_1(k) &= f_1(x_1 \mid k)\, Q_1(k), \\ F_{j+1}(k) &= f_{j+1}(x_{j+1} \mid k) \sum_{l=1}^{K} Q_{j+1}(k \mid l)\, F_j(l). \end{aligned}$$

The cost of computing all forward probabilities is generally $\mathcal{O}(pK^2)$. This can be reduced to $\mathcal{O}(pK)$ using the symmetries in (S7).

Proposition 8. *In the special case of the Li-and-Stephens model of phased haplotypes, the forward probabilities in Algorithm 4 can be computed recursively as follows:*

$$\begin{aligned} F_1(k) &= f_1(x_1 \mid k)\, a_{1,k}, \\ F_{j+1}(k) &= f_{j+1}(x_{j+1} \mid k) \left[a_{j+1,k} \sum_{l=1}^{K} F_j(l) + b_{j+1} F_j(k) \right]. \end{aligned}$$

Above, the sum only needs be computed once because it does not depend on the value k. Consequently, we can implement the first part of Algorithm S2 A 2 at cost $\mathcal{O}(pK)$.

3. *Unphased genotypes*

We can also implement Algorithm 4 efficiently for unphased genotypes. For $j \in \{2, \dots, p\}$, the forward probabilities are defined as:

$$\begin{aligned} F_1(\{k^a, k^b\}) &= f_1(x_1 \mid \{k^a, k^b\})\, \bar{Q}_1(\{k^a, k^b\}), \\ F_{j+1}(\{k^a, k^b\}) &= f_{j+1}(x_{j+1} \mid \{k^a, k^b\}) \sum_{\{l^a, l^b\}} \bar{Q}_{j+1}(\{k^a, k^b\} \mid \{l^a, l^b\})\, F_j(\{l^a, l^b\}). \end{aligned}$$

This computation generally costs $\mathcal{O}(pK^4)$ because the number of discrete states in the latent Markov chain is quadratic in the number of haplotype motifs.

Proposition 9. *In the special case of the Li-and-Stephens model of unphased genotypes, the forward probabilities in Algorithm 4 are given by the following recursive formula:*

$$\begin{aligned}
\frac{F_1(\{k^a, k^b\})}{f_1(x_1 \mid \{k^a, k^b\})} &= a_{1,k^a} a_{1,k^b} (2 - \delta_{k^a,k^b}), \\
\frac{F_{j+1}(\{k^a, k^b\})}{f_{j+1}(x_{j+1} \mid \{k^a, k^b\})} &= (b_{j+1})^2 F_j(\{k^a, k^b\}) + (2 - \delta_{k^a,k^b}) \left[a_{j+1,k^a} a_{j+1,k^b} \sum_{\{l^a,l^b\}} F_j(\{l^a, l^b\}) \right] \\
&\quad + b_{j+1} \frac{a_{j+1,k^a}}{1 + \delta_{k^a,k^b}} \left[F_j(\{k^b, k^b\}) + \sum_{l=1}^{K} F_j(\{l, k^b\}) \right] \\
&\quad + b_{j+1} \frac{a_{j+1,k^b}}{1 + \delta_{k^a,k^b}} \left[F_j(\{k^a, k^a\}) + \sum_{l=1}^{K} F_j(\{l, k^a\}) \right].
\end{aligned}$$

This gives us a $\mathcal{O}(pK^2)$ algorithm because the above sums can be efficiently pre-computed.

S3. TECHNICAL DETAILS

A. Schematic

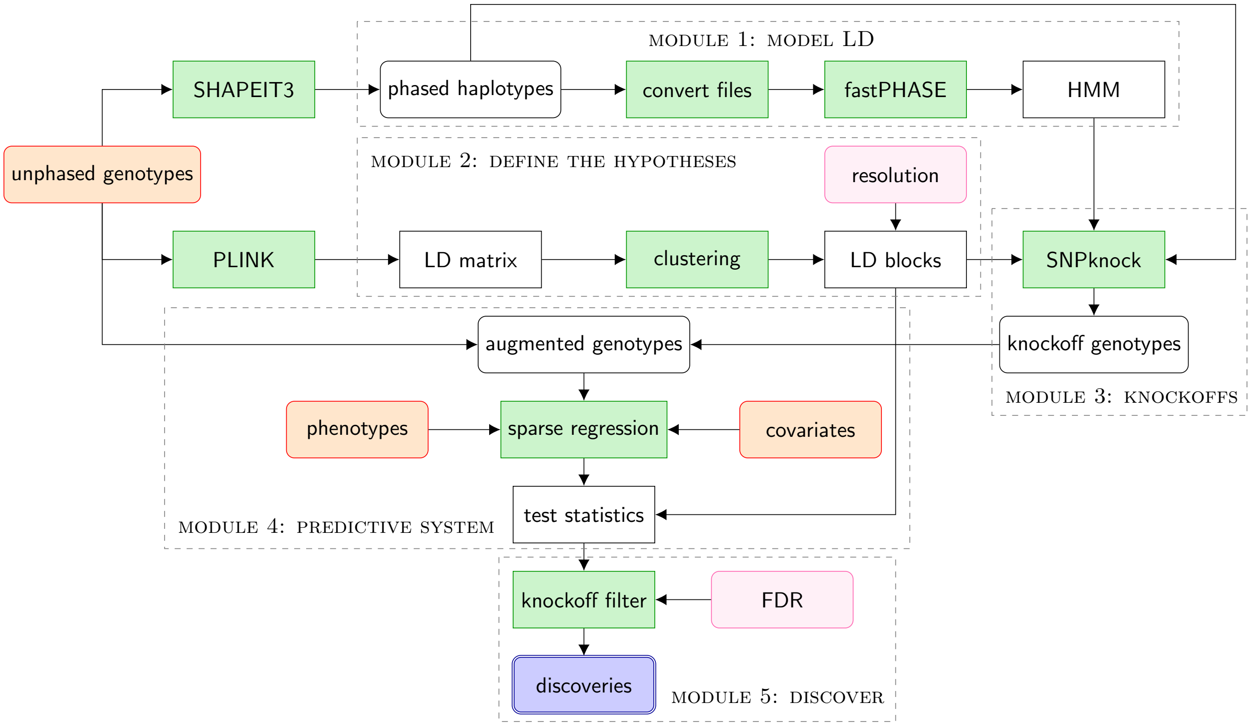


FIG. S1. Schematic of the *KnockoffZoom* method for the analysis of GWAS data. The inputs are the unphased genotypes (or the phased haplotypes, if available), the phenotypes, and any relevant covariates (e.g., age, sex, other demographics, principal components). The user chooses the resolution at which our method performs genome-wide fine-mapping and the nominal FDR level. In the first module, an HMM is fit using the phased haplotypes using fastPHASE. Meanwhile, the typed loci are assigned to blocks that represent our units of inference (second module), based on the observed LD, the physical locus positions and the resolution. Then, the estimated HMM and the partition of the variants are used by our software (SNPknock) to generate the knockoffs (third module). In the fourth module, a test statistic for each block is computed by a machine-learning system that predicts the phenotype and estimates the feature importance of genotypes and knockoffs, without knowing which one is which. The knockoffs serve as negative controls, allowing the filter to calibrate the statistics for FDR control (fifth module).

B. Computational and memory resources

The computation time required by *KnockoffZoom* for the data analysis in this paper is reported in Table S1, for each operation described in the flowchart of Figure S1. The entire procedure, starting from phased haplotypes, takes about 12 days—this is less than the time needed for phasing.[6] In principle, our procedure can be applied to unphased genotypes, although this is not recommended if the dataset is very large, for the knockoff generation would be slower. If we want to analyze different phenotypes, only the fourth and fifth modules need to be repeated (the latter requires negligible resources). The analysis of a new trait for the same individuals in the UK Biobank takes us less than 12 hours with an efficient implementation of the fourth module. As a comparison, BOLT-LMM takes between a few days and a week, depending on the phenotype. These are rough upper bounds because the exact time depends on the computing hardware and other factors.

Module	Operation	Time (h)	Memory (GB)	Machines	Cores
1	convert haplotypes	30	50	22	1
1	fit HMM	144	50	22	1
2	compute LD matrix	2	50	22	1
2	hierarchical clustering	8	200	22	1
3	generate knockoffs	72	30	22	1
3	combine augmented data	8	60	1	1
4	memory mapping	12	60	1	1
4	sparse regression	12	20	1	10

TABLE S1. Computation time and resources for each module of *KnockoffZoom*, for the analysis of the phenotype *height* in the UK Biobank (591k SNPs and 350k individuals).

The resource requirements of our method are also summarized in Table S1. The computations in modules 1–3 are divided between 22 machines, one for each autosome (our estimates refer to the longest one). The memory footprint is low, except for clustering, which requires about 200GB. If this becomes limiting, one can define the LD blocks differently, e.g., locally and concatenating the results. By comparison, BOLT-LMM requires approximately 100 GB with the same data.

We can easily analyzed the theoretical complexity of the first and third modules. In order to efficiently fit an HMM for the genotypes, the haplotypes are phased using SHAPEIT3;[6] then fastPHASE[4] is applied to the inferred haplotypes. The computational complexity of SHAPEIT3 and fastPHASE are $\mathcal{O}(pn \log n)$ and $\mathcal{O}(pnK)$, respectively, where p is the number of variants, n is the number of individuals and K is the number of haplotype motifs in the HMM. Our analytical calculations for the Li-and-Stephens HMM have reduced the cost of the third module to $\mathcal{O}(pnK)$. If the phased haplotypes are not available, a modified version of our algorithm has cost $\mathcal{O}(pnK^2)$. The second and fourth modules are more difficult to analyze theoretically, as different choices of tools are available to cluster the variants and compute the test statistics. For the latter, we rely on very fast and memory-efficient implementations of sparse linear and logistic regression.[7] The fifth and final module of *KnockoffZoom* is computationally negligible.

C. Computing the test statistics

The feature importance measures that we have adopted to define the test statistics of *KnockoffZoom* are based on sparse multivariate linear and logistic regression, since Bayesian and penalized regression approaches are currently the state-of-the-art for predicting complex traits.[8,9] However, our methodology could easily incorporate other scalable machine learning tools, like SVMs, random forests, and deep learning.[10–12] In principle, the test statistics of *KnockoffZoom* could also be computed using cheaper single-marker methods, e.g., by contrasting the univariate regression p-values for the original genotypes and the knockoffs; however, we do not recommend this for two reasons. First, our method is more powerful if used in combination with a multi-marker predictive model that explains a larger fraction of the variance in the phenotypes. Second, multivariate methods are less susceptible to the confounding effect of population structure.[13] Therefore, they improve our robustness against possible violations of the modeling assumptions, i.e., the ability of the HMM to correctly describe the real distribution of the genotypes.

In this paper, we use the R packages `bigstatsr` and `bigsnpr`[7] to fit a sparse generalized linear model the trait given the augmented matrix of explanatory variables $[Z, X, \tilde{X}] \in \mathbb{R}^{n\times(m+2p)}$. Here, Z is a matrix containing the covariates for all individuals, while X and $\tilde{X}$ indicate the observed genotypes and their knockoff copies, respectively. The regularization penalty is tuned using a modified form of 10-fold cross-validation. The regression coefficients for each variable are averaged over the 10 folds, finally obtaining an estimate of the quantity $\hat{\beta}_j(\lambda_{\text{CV}})$ from the Online Methods B. Since the knockoff-augmented genotype data is stored on an hard-drive as a memory-mapped file, the algorithm is memory efficient.[14] In addition, if the knockoffs in a group g are known in advance to have very little power as negative controls, e.g., $\max_{j\in G} r^2(X_j, \tilde{X}_j) > 0.99$, the corresponding hypothesis is automatically discarded by setting $W_g = 0$. This operation is independent of the phenotypes and preserves the symmetry required to control the FDR[15], but it may improve power.

D. Implementation of the LMM oracle

The FDR-controlling oracle used throughout our simulations is implemented as follows. Significant loci are clumped by applying the PLINK algorithm to the LMM p-values, over a discrete grid of possible values for the primary significance threshold. We use a wide logarithmic scale: 5×10^{-9}, 5×10^{-8}, 5×10^{-7}, 5×10^{-6}, 10^{-5}, 2×10^{-5}, 5×10^{-5}, 10^{-4}, 2×10^{-4}, 5×10^{-4}, 10^{-3}. The number of discoveries and the true FDP are computed for each threshold and experiment. To mitigate the discretization, we interpolate the results between the grid points. Then, we count the true discoveries corresponding to the most liberal threshold that controls the FDP (or the FDR, if the experiments are replicated multiple times with the same parameters).

S4. NUMERICAL SIMULATIONS

A. Goodness-of-fit of the HMM

We fit the HMM of Section S2 B to the phased haplotypes for the 350k individuals in the UK Biobank retained in our analysis. The parameters of 22 separate models are estimated applying fastPHASE separately within each autosome. The flexibility of the HMM is controlled by the number K of haplotype motifs, which is important for the performance of *KnockoffZoom*. If K is too small, the knockoffs may not have the right LD structure to serve as negative controls, resulting in an excess of false discoveries; if K is too large, over-fitting may render our procedure overly conservative and reduce power. In order to choose a good value of K, we consider a wide range of alternatives and evaluate the goodness-of-fit of each model in terms of its ability to correctly predict missing values in a hold-out sample. For this purpose we divide the individuals into a training set of size 349,119 and a validation set of size 10,000; then, we mask 50% of the haplotypes from the second set and use the fitted HMMs to reconstruct their value given the observed data. The goodness-of-fit is thus measured in terms of imputation error: lower values indicate a more accurate model. The results corresponding to chromosome 22 are shown in Table S2. Even though $K = 100$ seems optimal according to this metric, little improvement is observed above 50. Therefore, we choose $K = 50$ to generate knockoffs for the rest of the analysis, in the interest of computational speed. Finally, we verify that the goodness-of-fit of this HMM with $K = 50$ is significantly better than that of a multivariate Gaussian approximation of the genotype distribution,[15] with parameters estimated on all 359k samples, which has imputation error equal to 5.52%.

K	1	2	5	10	15	20	30	50	75	100
Imputation error (%)	14.59	10.23	6.74	5.57	5.04	4.77	4.4	4.01	3.75	3.71

TABLE S2. Out-of-sample imputation error of our HMM for haplotypes missing at random, as a function of the number K of motifs in the model.

B. Resolution and locus partitions

We apply *KnockoffZoom* at 7 levels of resolution. Each level corresponds to a specific partition of the 591k loci into contiguous LD blocks, as summarized in Table S3. At high resolution, the intrinsic difficulty of fine-mapping is reflected by the greater similarity between the original genotypes and their knockoff copies. In this setting, we need many samples or large effect sizes to distinguish a statistically significant contrast between the predictive power of genotypes and knockoffs. Conversely, knockoffs and genotypes have a smaller r^2 at lower resolutions.

Resolution	Number of blocks	Mean block width (Mb)	Mean block size (# SNPs)	Mean knockoff r^2
100%	591513	0.000	1.00	0.739
75%	443636	0.001	1.33	0.732
50%	295754	0.004	2.00	0.722
20%	118300	0.018	5.00	0.686
10%	59151	0.042	10.00	0.606
5%	29576	0.088	20.00	0.458
2%	11831	0.226	50.0	0.231

TABLE S3. Summary of the genome partitions at different resolutions. The last column indicates the r^2 between genotypes and knockoffs at each resolution.

C. Genetic architecture of the synthetic phenotypes

We generate synthetic phenotypes with a variety of genetic architectures in order to test the performance of our method in controlled environments, as summarized in Table S4.

Name	Number of causal regions	Causal variants per region	Total number of causal variants	Width of causal regions (Mb)	Figures
1	500	5	2500	0.1	2–4, S2, S4–S6, S8
2	500	1	500	0.1	6
3	100	1	100	0.1	S7
4	100	5	500	0.1	S7
5	1000	1	1000	0.1	S3, S5, S7
6	1000	2	2000	0.1	S7
7	1000	5	5000	0.1	S7
8	1000	10	10000	0.1	S7

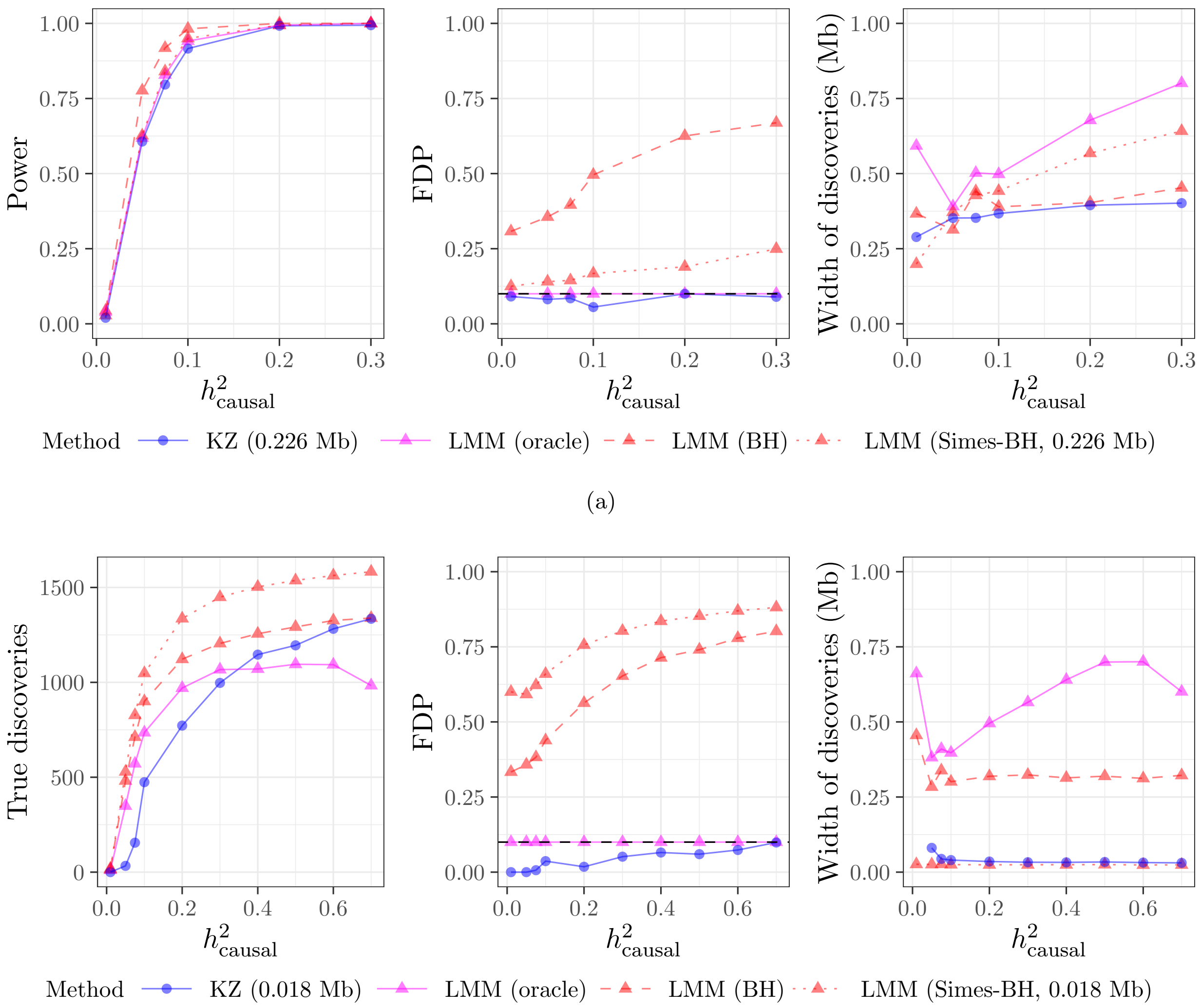
TABLE S4. Genetic architectures used in the numerical simulations. The first architecture is used in the main numerical experiments in the paper, while the simulated case-control study is based on the second one. The other architectures are considered in the additional experiments presented in this supplement.

D. The difficulty of controlling the FDR

Targeting the FDR in a GWAS is difficult; in particular, it is challenging to control it over distinct (conditional) discoveries using the LMM (marginal) p-values. To illustrate the challenges, we consider two variations of the Benjamini-Hochberg (BH) procedure.[16] First, we apply the BH correction to the BOLT-LMM p-values and then report significant discoveries with the usual PLINK clumping procedure. Second, we use BOLT-LMM to test pre-determined hypotheses defined over the LD blocks used by *KnockoffZoom*. For this purpose, we summarize the p-values corresponding to SNPs in the same block into a single number, a Simes p-value,[17] that we provide to the BH procedure without further clumping. For example, we consider two alternative resolutions: LD blocks that are 0.226 Mb or 0.018 Mb wide, on average.

The observed outcome is an excess of false positives, regardless of how we count distinct discoveries; see Figure S2. With the first approach, the FDR is inflated because the BH procedure counts the discoveries *pre-clumping*, while the FDR is evaluated on findings that are defined differently, *post-clumping*.[18] Moreover, these p-values are only designed for marginal hypotheses, while we are interested in distinct (conditional) findings.[19] With the second approach, the FDR would be controlled if the p-values for loci in different groups were correct and independent. However, this is not the case, even when the blocks are large, because some inter-block LD remains. These examples are not exhaustive, but they illustrate the fundamental problem that the LMM p-values are not calibrated for conditional hypotheses. We have shown in Section II C 2 that this limitation may invalidate even the typical FWER claims, whenever one attempts to distinguish between discoveries involving loci that are not very distant.

A more sophisticated version of the BH procedure has been recently proposed in combination with clumping,[18] although it does not directly address conditional hypotheses. Alternatively, an earlier solution is based on the distributions of scan statistics.[19] We are not aware of whether this has been tested, on simulated or real data. A different approach based on penalized regression[20] has shown promise, but it relies on stringent modeling assumptions.



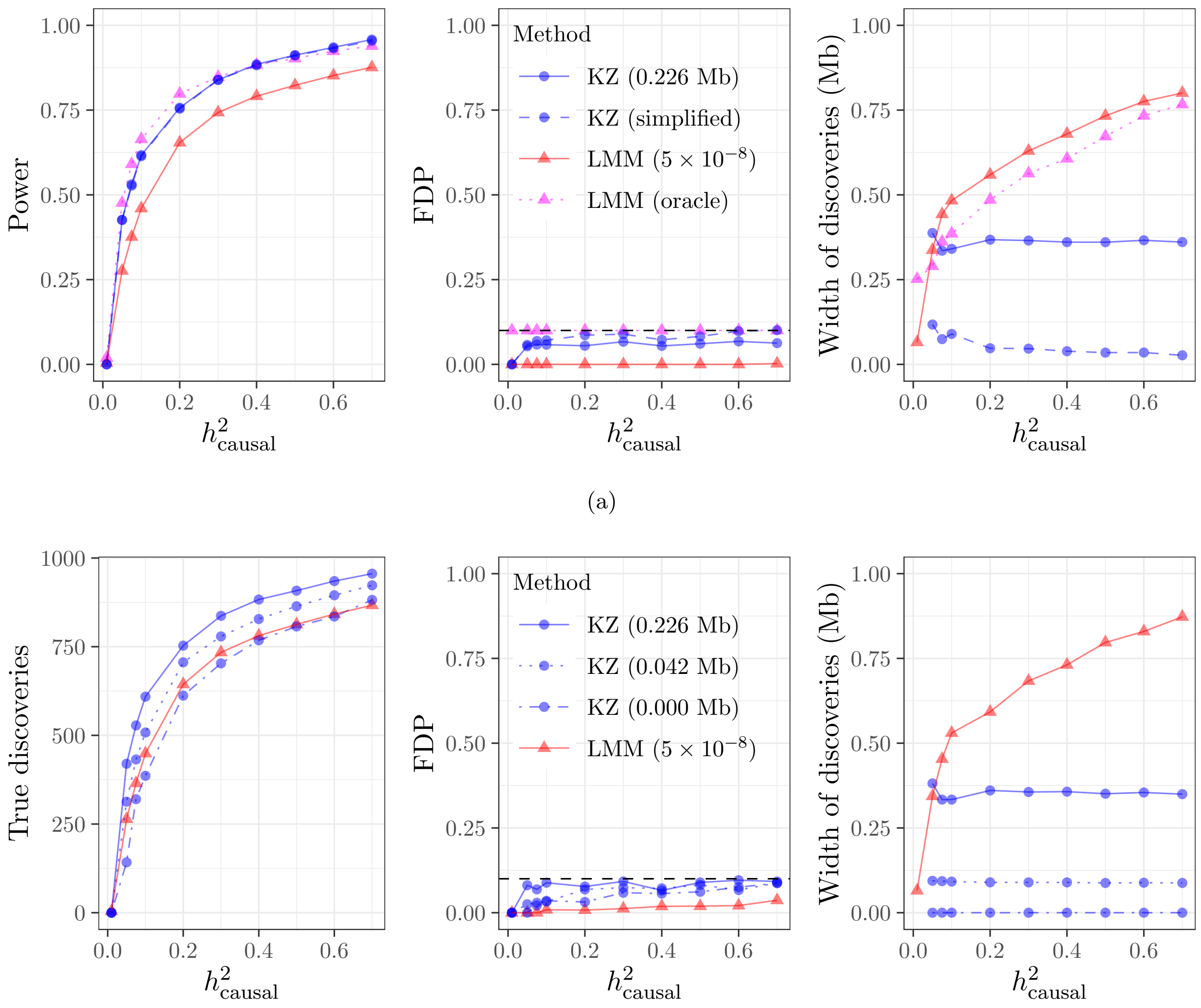
(a)

(b)

FIG. S2. Performance of *KnockoffZoom* and two alternative LMM-based heuristics for FDR control, for the same experiments as in Figure 3. In (b) we do not simplify the *KnockoffZoom* findings at multiple resolutions; instead, we only report those at a fixed resolution, for comparison with the Simes-BH method.

E. Locus discovery

We complement in Figure S3 the simulations in Figure 3 by considering the fifth architecture from Table S4, in which all causal SNPs are well-separated. We observe in Figure S3 (b) that, if the signals are weak, it is difficult to reject the high-resolution conditional hypotheses. If the signals are sufficiently strong, we separate nearby casual variants and make additional distinct discoveries.



(a)

(b)

FIG. S3. Performance of *KnockoffZoom* and BOLT-LMM for a simulated trait with genetic architecture 5 from Table S4. (b): *KnockoffZoom* at different levels of resolutions. Other details as in Figure 3.

F. Fine-mapping

We provide further details about the fine-mapping simulations discussed in Section II C 3 and present additional results obtained with an alternative genetic architecture.

We calibrate the two-step procedure involving BOLT-LMM followed by SUSIE or CAVIAR to target a nominal error rate that is comparable to the FDR of *KnockoffZoom*. Regarding SUSIE, we simply apply it with nominal coverage parameter equal to 90%, so that at most 10% of its reported findings are expected to be false positives. In the case of CAVIAR, the solution is similar but it requires a more careful explanation. Under certain modeling assumptions, CAVIAR is designed to control the probability that any causal variants are erroneously discarded from the input candidate set[21] (but it does not tell apart distinct causal effects, which is a major limitation[22]). Therefore, assuming that the input always contains at least one causal SNP (as we ensure in our simulations by applying BOLT-LMM with aggressive clumping), CAVIAR indirectly targets, within the two-step procedure, a notion of FDR similar to that of SUSIE and *KnockoffZoom*. In particular, setting the control parameter of CAVIAR equal to 90%, we can expect that at most 10% of the reported findings do not contain at least one causal variant.

A downside of the two-step paradigm of fine-mapping is that it may sometimes introduce selection bias, as shown in Figure S4. The biased results refer to SUSIE applied exactly on the clumps reported by BOLT-LMM, without including nearby unselected SNPs.

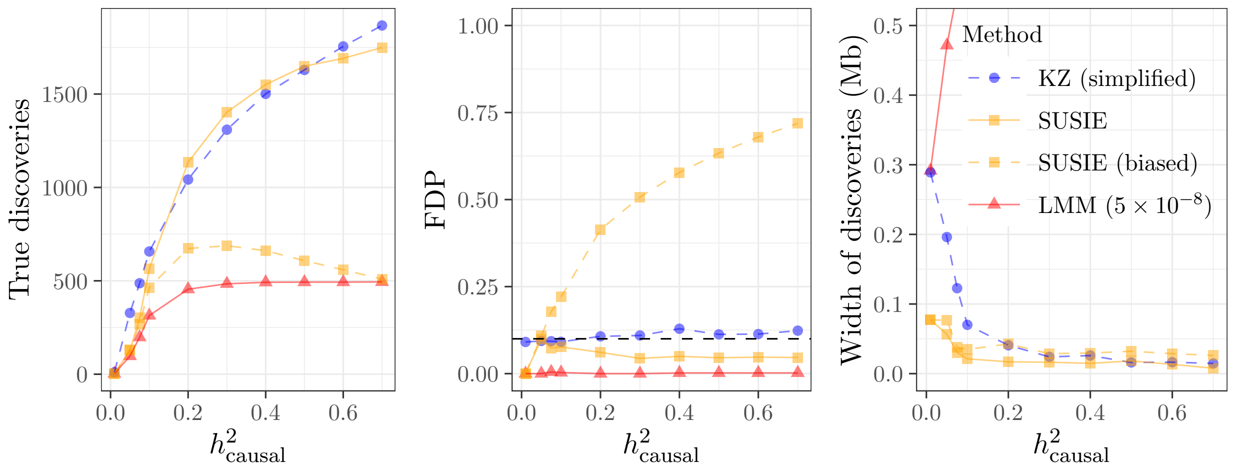


FIG. S4. Fine-mapping with BOLT-LMM followed by SUSIE, with and without mitigating the selection bias. Other details as in Figure 4.

We complement the results in Figure 4 by considering the fifth architecture from Table S4, as in Figure S3. Here, *KnockoffZoom* is more powerful than SUSIE. Note that the output of SUSIE needs to be simplified because it occasionally reports overlapping subsets of SNPs; therefore, we consolidate those that are not disjoint and retain the smallest if one is included in the other.

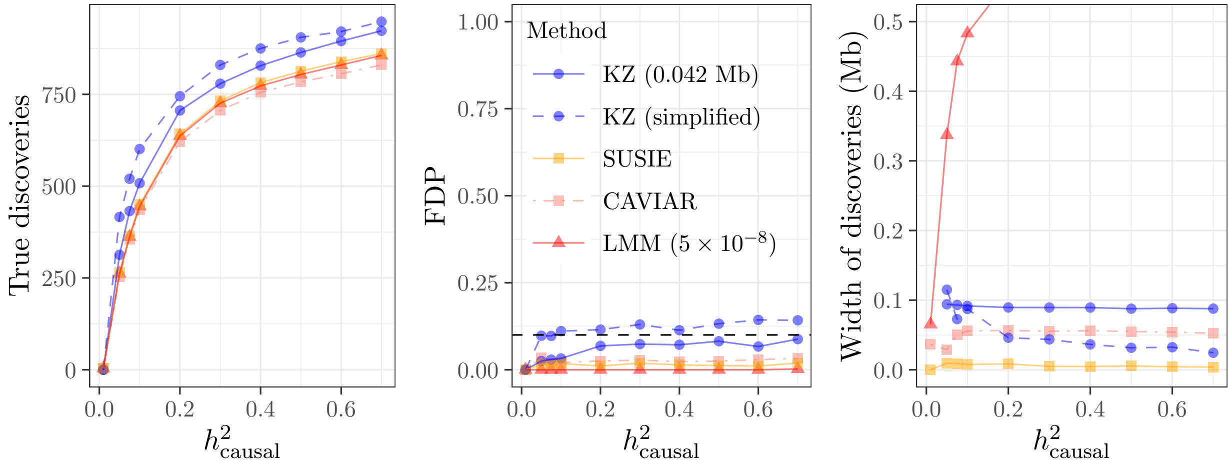
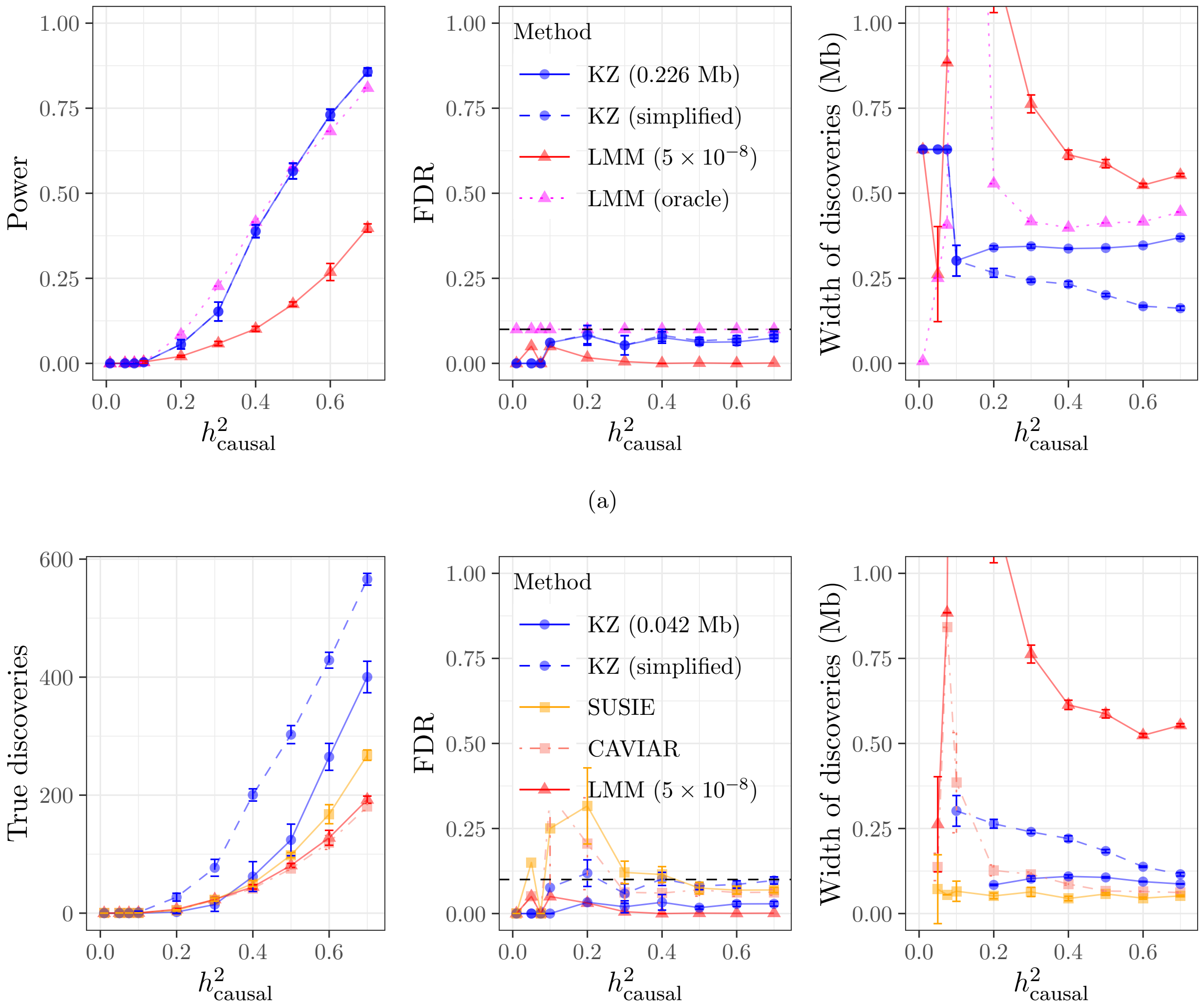


FIG. S5. Fine-mapping performance of *KnockoffZoom* and BOLT-LMM followed by CAVIAR or SUSIE. Simulated trait as in Figure S3. Other details as in Figure 4.

G. Repeated simulations with smaller sample size

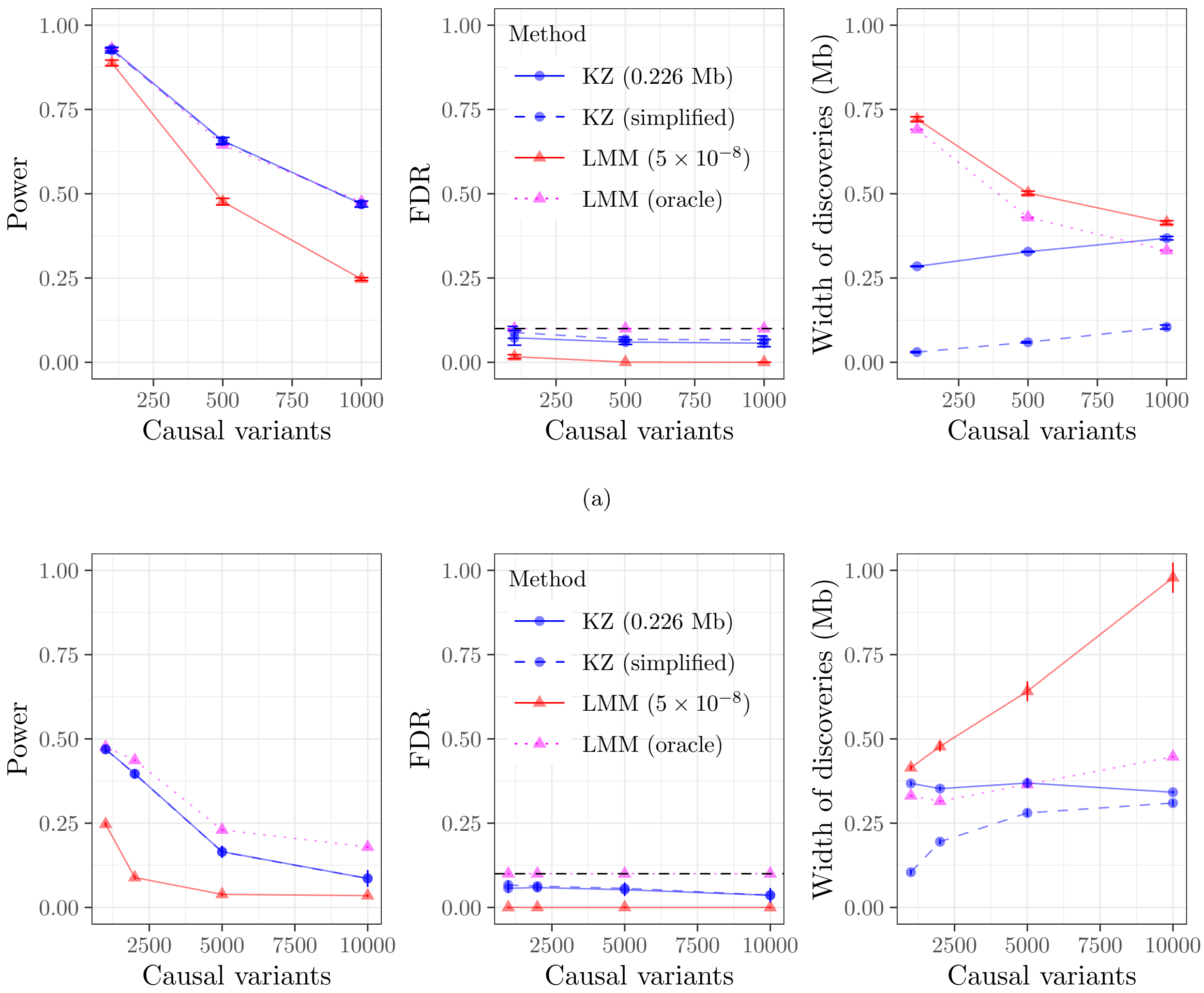
We investigate the average behaviour of *KnockoffZoom* and its alternatives on multiple independent realizations of the genotype data and the simulated phenotypes. We divide the available genotype observations into 10 disjoint subsets, each containing 30k individuals, and analyze them separately after generating artificial phenotypes for each. The FDR is estimated by averaging the FDP over the 10 repetitions. The results are in Figures S6 and S7.



(a)

(b)

FIG. S6. Performance of *KnockoffZoom* and BOLT-LMM for simulated phenotypes, repeating the experiments 10 times on disjoint subsets of the data, each including 30k individuals. The error bars indicate 95% confidence intervals for the mean quantities, estimated from 10 independent replications of the trait given the same genotypes. The other details in (a) and (b) are as in Figures 3 and 4, respectively.



(a)

(b)

FIG. S7. Performance of *KnockoffZoom* and BOLT-LMM for a simulated trait with different genetic architectures from Table S4, as a function of the number of causal variants. The heritability of the trait is $h^2_{\text{causal}} = 0.5$. (a): architectures 3,4,5; (b): architectures 5,6,7,8. Other details as in Figure S6 (a).

H. Assessing the individual significance of each discovery

KnockoffZoom controls the FDR so that at most a pre-specified fraction q of the discoveries are false positives, on average. In addition to this global guarantee, we can quantify, to some extent, the statistical significance of the individual findings. We have not discussed this in the paper in the interest of space and because the underlying theory is not fully developed. Nonetheless, the main idea is intuitive. Keep in mind that the basic ingredient of the knockoff filter[23] is a suitable estimate of the false discovery proportion: FDP(t), i.e., the fraction of false discoveries if we reject all hypotheses in $\{g : W_g \geq t\}$. Note that the test statistics W_g are defined in Online Methods B. Additional information can be extracted from the test statistics by estimating a local version of the FDR[24], e.g., the fraction of false discoveries in $\{g : t - \Delta t_1 \leq W_g \leq t + \Delta t_2\}$, for some choice of Δt_1 and Δt_2. For this purpose, we propose the following estimator:

$$\widehat{\text{Fdp}}(t, \Delta t_1, \Delta t_2) = \frac{c + |\{g : -(t + \Delta t_2) \leq W_g \leq \min(0, -t + \Delta t_1)\}|}{\max(1, |\{g : \max(0, t - \Delta t_1) \leq W_g \leq t + \Delta t_2\}|)}, \tag{S10}$$

with either $c = 0$ (more liberal) or $c = 1$ (more conservative). The usual estimate of the FDP computed by the knockoff filter[23] is recovered if $\Delta t_1 = 0$ and $\Delta t_2 \to \infty$:

$$\widehat{\text{FDP}}(t) = \frac{c + |\{g : W_g \leq -t\}|}{\max(1, |\{g : W_g \geq t\}|)},$$

We expect that the findings whose test statistics fall in regions of low $\widehat{\text{Fdp}}$ are less likely to be false positives. The choices of Δt_1 and Δt_2 determine the accuracy of our estimates: if the span is smaller, the local FDR encodes information at higher resolution; however, our estimate is noisier because it must rely on fewer statistics. We compute the estimate in (S10) in simulations (Figure S8) and on real data (Figure S9), setting $\Delta t_1 = 100 = \Delta t_2$ and $c = 0$ (by contrast, we use $c = 1$ in the knockoff filter). This approach is reasonable if *KnockoffZoom* reports sufficiently many discoveries; otherwise, the estimated local FDP may have very high variance.

The simulation in Figure S8 is carried out on an artificial phenotype with the first genetic architecture in Table S4 and heritability $h^2_{\text{causal}} = 0.7$. We plot the FDP as a function of the number of selected variants, based on the ordering defined by the *KnockoffZoom* test statistics.

We can draw two interesting observations from these results. First, the estimated cumulative FDP tracks the curve corresponding to the true FDP closely, which is consistent with the fact that the FDP of *KnockoffZoom* is almost always below the nominal FDR level throughout our simulations (Section II C), even though the theoretical guarantee refers to the average behavior. The low variance in the estimated FDP may be explained by the size of the dataset and the large number of discoveries. Second, the estimated local FDP also approximates the corresponding true quantity quite precisely, especially for the statistics above the rejection threshold. This may be very valuable in practice, because it provides us with a good educated guess of which discoveries are likely to be false positives. Whether we can rigorously state more precise results is an interesting question that we plan to explore deeper in the future.

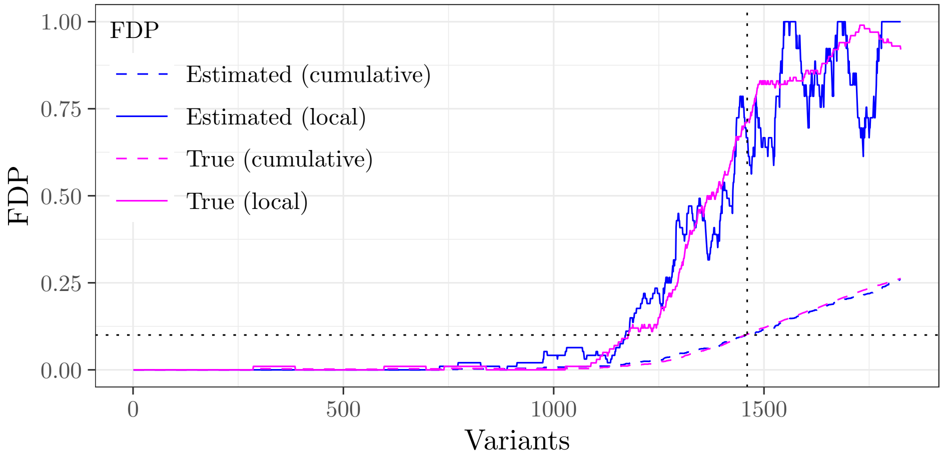


FIG. S8. Estimated and true proportion of false discoveries obtained with *KnockoffZoom* at medium resolution (0.018 Mb), for a simulated trait with 2500 causal variants. The dotted vertical line indicates the adaptive significance threshold computed by the knockoff filter at the nominal FDR significance of 0.1. This corresponds exactly to the last crossing of the 0.1 level (dotted horizontal line) by the estimated cumulative FDP curve. The local (or estimated) FDP is computed by looking at 100 statistics within a rolling window.

S5. DATA ANALYSIS

A. Phenotype definition in the UK Biobank

We have analyzed the phenotypes in the UK Biobank data defined in Table S5.

Name	Description	Number of cases	UK Biobank Fields	UK Biobank Codes
height	standing height	continuous	50-0.0	
bmi	body mass index	continuous	21001-0.0	
sbp	systolic blood pressure	continuous	4080-0.0, 4080-0.1	
platelet	platelet count	continuous	30080-0.0	
cvd	cardiovascular disease	116454	20002-0.0–20002-0.32	1065, 1066, 1067, 1068, 1081, 1082, 1083, 1425, 1473, 1493
diabetes	diabetes	14848	20002-0.0–20002-0.32	1220
hypothyroidism	hypothyroidism	17985	20002-0.0–20002-0.32	1226
respiratory	respiratory disease	51625	20002-0.0–20002-0.32	1111, 1112, 1113, 1114, 1115, 1117, 1413, 1414, 1415, 1594
glaucoma	glaucoma	2656	4689-0.0–4689-2.0	age $>$ 30

TABLE S5. Phenotype definition in the UK Biobank. The numbers of disease cases refer to the subset of unrelated British individuals that passed our quality control.

B. Number of discoveries

Table S6 shows that the inclusion of the principal components has little effect on our analysis, which confirms that we are not significantly affected by confounding due to population structure, at least within this cohort of unrelated British individuals.

Phenotype	Resolution						
	0.226 Mb	0.088 Mb	0.042 Mb	0.018 Mb	0.004 Mb	0.001 Mb	0.000 Mb
height	3284 – 3252	1976 – 2007	823 – 785	388 – 391	336 – 335	170 – 214	173 – 167
bmi	1804 – 1808	555 – 471	60 – 46	33 – 33	24 – 24	0 – 21	15 – 17
platelet	1460 – 1479	890 – 880	408 – 413	276 – 236	161 – 200	181 – 156	143 – 146
sbp	722 – 745	297 – 322	95 – 129	0 – 0	0 – 0	0 – 0	0 – 0
cvd	514 – 475	182 – 164	51 – 51	0 – 0	0 – 0	0 – 0	0 – 0
hypothyroidism	212 – 163	108 – 103	0 – 0	0 – 0	0 – 0	0 – 0	21 – 21
respiratory	176 – 183	65 – 60	41 – 12	13 – 13	14 – 14	12 – 12	0 – 0
diabetes	50 – 48	33 – 33	21 – 18	10 – 10	11 – 12	10 – 10	0 – 0
glaucoma	0 – 0	0 – 0	0 – 0	0 – 0	0 – 0	0 – 0	0 – 0

TABLE S6. Number of distinct findings made with *KnockoffZoom* at different resolutions, with and without using the top 5 principal components as covariates. The findings obtained with the principal components correspond to those reported in Table I.

Table S7 quantifies the overlap between our low-resolution discoveries (including the principal components) and those obtained with BOLT-LMM applied to the same data. We say that findings made by different methods are overlapping if they are within 0.1 Mb of each other. These results confirm that our method makes many new findings, in addition to refining those reported by BOLT-LMM.

Phenotype	Method	Discoveries	Overlapping	Distinct
height	*KnockoffZoom*	3284	2246	1038
	BOLT-LMM	1685	1679	6
bmi	*KnockoffZoom*	1804	627	1177
	BOLT-LMM	389	378	11
platelet	*KnockoffZoom*	1460	848	612
	BOLT-LMM	723	709	14
sbp	*KnockoffZoom*	722	243	479
	BOLT-LMM	197	188	9
cvd	*KnockoffZoom*	514	188	326
	BOLT-LMM	156	144	12
hypothyroidism	*KnockoffZoom*	212	96	116
	BOLT-LMM	96	91	5
respiratory	*KnockoffZoom*	176	59	117
	BOLT-LMM	63	59	4
diabetes	*KnockoffZoom*	50	40	10
	BOLT-LMM	47	43	4
glaucoma	*KnockoffZoom*	0	0	0
	BOLT-LMM	5	0	5

TABLE S7. Overlap of the discoveries made with *KnockoffZoom* at low resolution (0.226 Mb) and BOLT-LMM (clumped without consolidation), using the same data. *KnockoffZoom* is applied as in Table I. For example, our method reports 3284 findings for *height*, 2246 of which overlap at least one of the 1685 clumps found by the LMM, while 1038 are distinct from those found by the LMM. Conversely, only 6 out of 1685 discoveries made by BOLT-LMM are not detected by *KnockoffZoom*.

C. Comparison with BOLT-LMM on a larger sample

Our discoveries using data from 350k unrelated British individuals are compared in Table S8 to those obtained in earlier work by applying BOLT-LMM to all 459,327 European subjects in the UK Biobank.[25] We say that findings made by different methods are overlapping if they are within 0.1 Mb of each other. The LMM makes fewer findings despite the larger sample size. This indicates that our method is more powerful, in addition to providing more interpretable results.

Phenotype	*KnockoffZoom* 350k unrelated British			BOLT-LMM 459k European			
		Overlapping with BOLT-LMM		(5×10^{-9})		(5×10^{-8})	
	Discoveries	(5×10^{-9})	(5×10^{-8})	Discoveries	Overlapping with *KnockoffZoom*	Discoveries	Overlapping with *KnockoffZoom*
height	3284	2339	2547	2056	2033	2464	2431
bmi	1804	778	967	504	493	697	672
platelet	1460	1025	1111	1016	988	1204	1155
sbp	722	401	461	440	371	568	452
cvd	514	203	257	192	173	257	229
hypothyroidism	212	111	137	118	112	143	134

TABLE S8. Comparison of the low-resolution (0.226 Mb) discoveries reported by *KnockoffZoom*, as in Table I, and those obtained by BOLT-LMM with a larger dataset. For example, BOLT-LMM reports 2056 discoveries for *height* at the significance level 5×10^{-9}, 2033 of which overlap with at least one of our 3284 findings, while 2339 of our findings overlap with at least one discovery made by BOLT-LMM.

D. Reproducibility (low resolution)

Phenotype	Method	# Discoveries		
		Total	Overlapping	Distinct
height	*KnockoffZoom*	121	56	65
	BOLT-LMM	54	49	5
platelet	*KnockoffZoom*	81	44	37
	BOLT-LMM	47	42	5

TABLE S9. Overlap of the discoveries made with *KnockoffZoom* (low-resolution) and BOLT-LMM (clumped without consolidation). Other details as in Table II. For example, we make 121 discoveries for *height*, 56 of which are within 0.1 Mb of at least one of the 54 clumps found by the LMM, while 65 are distinct. In this case, only 5 out of 54 discoveries made by BOLT-LMM are not detected by *KnockoffZoom*.

		30k unrelated British			459k European (BOLT-LMM)			
		Discoveries			Discoveries		Discoveries (consolidated)	
Phenotype	Method	#	Not replicated	Size (Mb)	Anticipated	Size (Mb)	Anticipated	Size (Mb)
height	*KnockoffZoom*	121	8 (6.6%)	0.308	398/2056 (19.4%)	0.789	92/807 (11.4%)	1.162
	LMM	54	0 (0.0%)	0.965	280/2056 (13.6%)	0.789	48/807 (5.9%)	1.162
	LMM (BH)	714	203 (28.4%)	0.379	1047/2056 (50.9%)	0.789	325/807 (40.3%)	1.162
platelet	*KnockoffZoom*	81	5 (6.2%)	0.319	235/1016 (23.1%)	0.805	65/525 (12.4%)	0.953
	LMM	47	0 (0.0%)	0.674	219/1016 (21.6%)	0.805	42/525 (8.0%)	0.953
	LMM (BH)	272	92 (33.8%)	0.433	422/1016 (41.5%)	0.805	136/525 (25.9%)	0.953

TABLE S10. Estimated low-resolution power of *KnockoffZoom* and BOLT-LMM. Other details as in Table S9. We say that a finding reported by BOLT-LMM on the larger dataset is anticipated by those in the smaller dataset if it is within 0.1 Mb of at least one of them. For example, among the 2056 unconsolidated discoveries reported by BOLT-LMM for *height* on the large dataset, 398 are anticipated by the findings of *KnockoffZoom* on the smaller dataset, while only 280 are anticipated by BOLT-LMM using the same data.

E. Reproducibility (high resolution)

Resolution	Platelet activation	Hemostasis	Blood coagulation	Wound healing	Response to wounding
0.226	1.8×10^{-08}	2.8×10^{-12}	3.9×10^{-12}	2.0×10^{-11}	5.6×10^{-13}
0.088	2.8×10^{-10}	3.1×10^{-20}	6.1×10^{-20}	1.2×10^{-19}	4.8×10^{-21}
0.042	1.0×10^{-08}	1.1×10^{-20}	3.6×10^{-20}	3.6×10^{-18}	6.8×10^{-17}
0.018	5.4×10^{-08}	3.0×10^{-21}	1.3×10^{-20}	4.0×10^{-18}	3.9×10^{-17}
0.004	2.8×10^{-09}	5.4×10^{-23}	3.4×10^{-22}	1.8×10^{-19}	4.7×10^{-17}
0.001	3.5×10^{-09}	3.0×10^{-23}	1.8×10^{-22}	3.1×10^{-20}	3.5×10^{-19}
0.000	2.6×10^{-09}	1.8×10^{-24}	1.4×10^{-23}	5.4×10^{-21}	1.8×10^{-17}

TABLE S11. Gene ontology enrichment analysis using GREAT[26] for the *KnockoffZoom* results on *platelet*. The uncorrected p-values for 5 relevant biological processes are shown at each resolution.

Consequence	Discoveries
nonsense	1
missense	26
synonymous	2
3-prime UTR	7
splice donor	1
intronic	74
500B downstream	2
2KB upstream	4
intergenic	26

TABLE S12. Most serious previously known consequence (dbSNP[27]) of each of the 143 variants discovered by *KnockoffZoom* at the highest resolution for the phenotype *platelet* in the UK Biobank.

Previous association	Discoveries
platelet	61
hemoglobin	4
red cells	2
squamous cell carcinoma	2
white cells	2
breast cancer	1
coronary artery disease	1
fatty liver disease	1
macular degeneration	1
menarche (age of onset)	1
obesity	1
none	66

TABLE S13. Most relevant previously reported associations for some of the 143 variants discovered by *KnockoffZoom* at the highest resolution for the phenotype *platelet* in the UK Biobank.

Chromosome	SNP	Position	Gene	Consequence
19	Affx-15656246	19765499	ATP13A1	missense
19	rs12983010	39229089	CAPN12	missense
1	rs140584594	110232983	GSTM1	missense
17	rs79007502	33880305	SLFN14	missense
2	rs76774368	160604514	MARCH7	missense
3	rs34095724	184099050	CHRD	missense

TABLE S14. Missense variants localized by *KnockoffZoom* at the highest resolution, for the phenotype *platelet* in the UK Biobank, that have not been reported before.

F. Assessing the individual significance of each discovery

We assess the individual significance of our discoveries on real data as discussed in Section S4 H. Figure S9 shows an example of the estimated local and global FDP as a function of the number of selected variants, based on the ordering defined by the *KnockoffZoom* test statistics. This information is included in the visualization of our discoveries in Section S5 G.

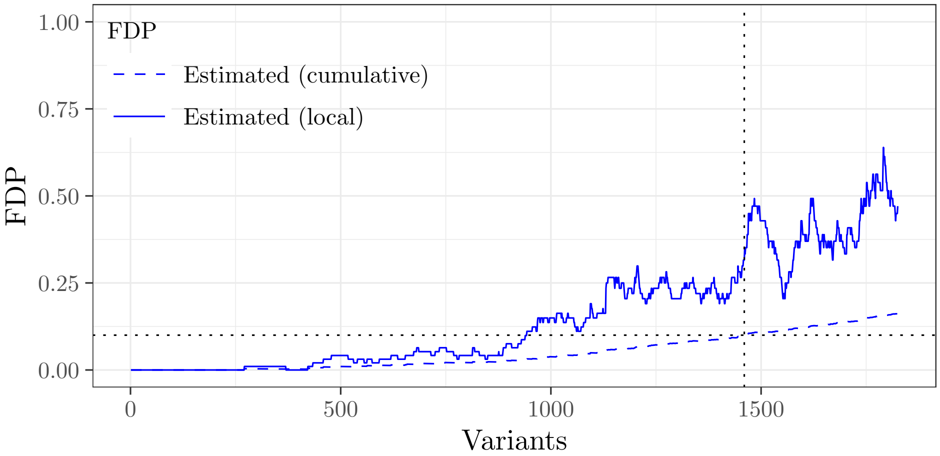


FIG. S9. Estimated proportion of false discoveries obtained with *KnockoffZoom* at low-resolution (0.226 Mb) for the phenotype *platelet* in the UK Biobank. Other details as in Figure S8.

G. Chicago plots

Figure 6 complements the example in Figures 1 and 5 by including an estimate of the local FDP computed as in Figure S9, for each level of resolution. Moreover, we also include the results obtained with BOLT-LMM for comparison.

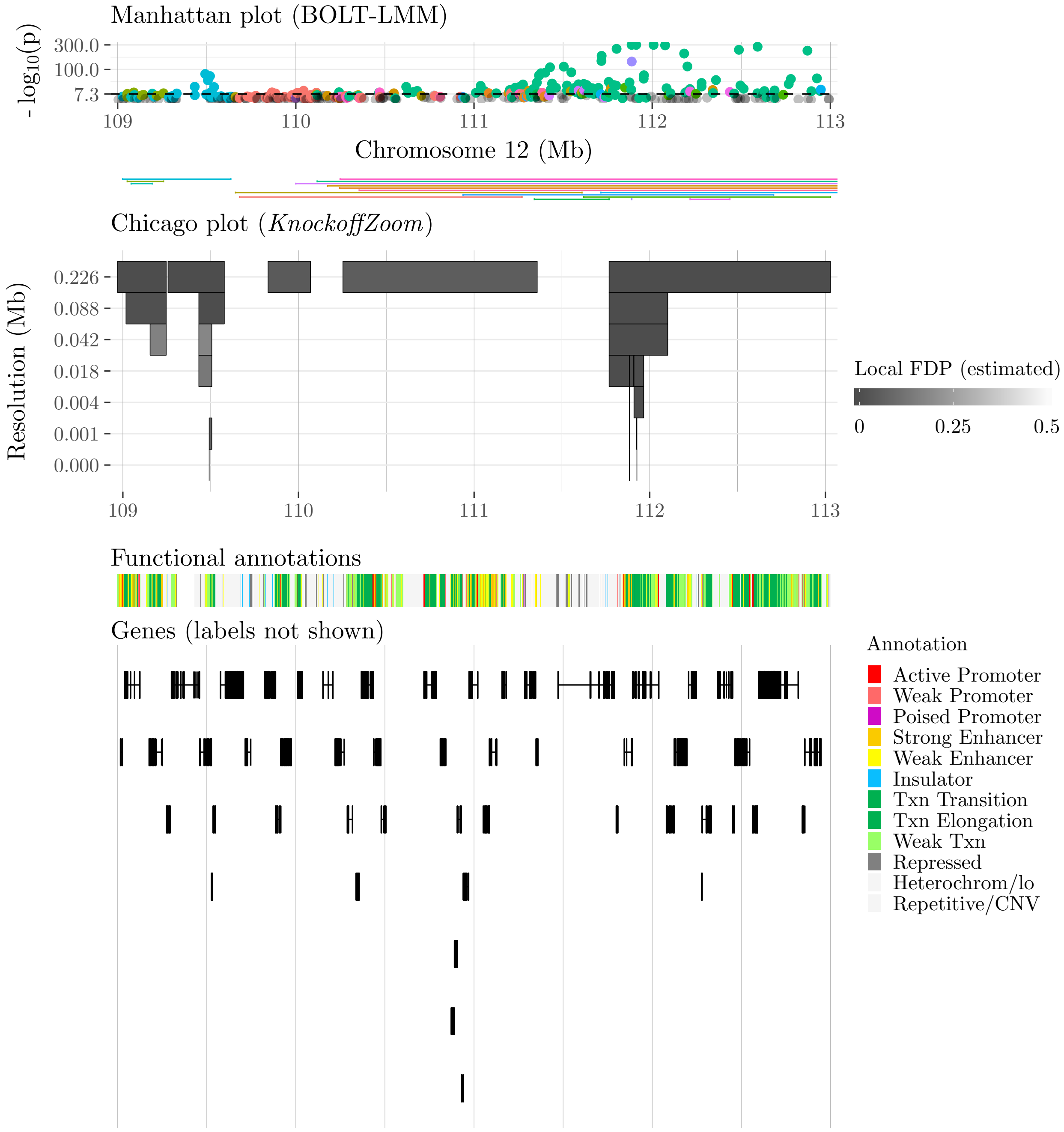


FIG. S10. Visualization of some discoveries made with *KnockoffZoom* and BOLT-LMM for *platelet* in the UK Biobank. Other details as in Figures 1 and 5. This Chicago plot is upside-down.

S6. MATHEMATICAL PROOFS

Proof of Proposition 2. Our algorithm implements the SCIP recipe,[15] which provably generates valid knockoffs, for the vector-valued random variables $Z^1, \ldots, Z^L$. Therefore, it suffices to show that (S3) gives a correct expression for $p(Z^g \mid Z^{-g}, \tilde{Z}^{1:(g-1)})$. To this end, we proceed by induction. Assuming that (S3) holds $\forall g \in \{1, \ldots, g'-1\}$, for some $g' > 1$, it follows that:

$$\begin{aligned}
p(Z^{g'} &| Z^{-g'}, \tilde{Z}^{1:(g'-1)}) \\
&\propto p(Z^{g'}, Z^{-g'}) p(\tilde{Z}^{1:(g'-1)} \mid Z^{g'}, Z^{-g'}) \\
&\propto Q_1^{g'}(Z_1^{g'} \mid Z_{m_{g'-1}}^{g'-1}) \left[\prod_{j=2}^{m_{g'}} Q_j^{g'}(Z_j^{g'} \mid Z_{j-1}^{g'}) \right] Q_1^{g'+1}(Z_1^{g'+1} \mid Z_{m'_g}^{g'}) p(\tilde{Z}^{g'-1} \mid Z^{g'}, Z^{-g'}) \\
&\propto Q_1^{g'}(Z_1^{g'} \mid Z_{m_{g'-1}}^{g'-1}) \left[\prod_{j=2}^{m_{g'}} Q_j^{g'}(Z_j^{g'} \mid Z_{j-1}^{g'}) \right] Q_1^{g'+1}(Z_1^{g'+1} \mid Z_{m'_g}^{g'}) \frac{Q_1^{g'}(Z_1^{g'} \mid \tilde{Z}_{m_{g'-1}}^{g'-1})}{\mathcal{N}_{g'-1}(Z_1^{g'})}.
\end{aligned}$$

Above, the proportionality holds across values of the vector $Z^{g'}$. In view of the normalization in (S4), we conclude that (S3) gives a correct expression for $p(Z^g \mid Z^{-g}, \tilde{Z}^{1:(g-1)})$ for all $g \in \{1, \ldots, g'\}$. Since the base case with $g = 1$ clearly holds, the proof is complete. □

Proof of Lemma 1. This follows immediately by replacing (S7) into (S8) and simplifying. □

Proof of Proposition 3. It suffices to prove (S6), since marginalizing over $(Z, \tilde{Z})$ implies that $(X, \tilde{X})_{\text{swap}(G;\mathcal{G})}$ has the same distribution as $(X, \tilde{X})$.[2] Conditioning on the latent variables yields:

$$\begin{aligned}
\mathbb{P}\Big[(X,\tilde{X}) &= (x, \tilde{x})_{\text{swap}(G;\mathcal{G})} \mid (Z, \tilde{Z}) = (z, \tilde{z})_{\text{swap}(G;\mathcal{G})}\Big] \\
&= \mathbb{P}\left[(X, \tilde{X}) = (x, \tilde{x})_{\text{swap}(G;\mathcal{G})} \middle| (Z, \tilde{Z}) = (z, \tilde{z})_{\text{swap}(G;\mathcal{G})} \right] \mathbb{P}\left[(Z, \tilde{Z}) = (z, \tilde{z})_{\text{swap}(G;\mathcal{G})} \right] \\
&= \mathbb{P}\left[(X, \tilde{X}) = (x, \tilde{x}) \middle| (Z, \tilde{Z}) = (z, \tilde{z}) \right] \mathbb{P}\left[(Z, \tilde{Z}) = (z, \tilde{z})_{\text{swap}(G;\mathcal{G})} \right] \\
&= \mathbb{P}\left[(X, \tilde{X}) = (x, \tilde{x}) \middle| (Z, \tilde{Z}) = (z, \tilde{z}) \right] \mathbb{P}\left[(Z, \tilde{Z}) = (z, \tilde{z}) \right].
\end{aligned}$$

Above, the first equality follows from the first line of Algorithm 3, the second from the conditional independence of the emission distributions in an HMM, and the third from Proposition 2). □

Proof of Proposition 4. The $\mathcal{N}$ function for the g-th group can be written as:

$$\begin{aligned}
\mathcal{N}_g(k) = \sum_{z_1^g, \ldots, z_{m_g-1}^g} & \frac{Q_1^g(z_1^g \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g \neq 1]} \, Q_1^g(z_1^g \mid z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(z_1^g)} \times \left[\prod_{j=2}^{m_g-1} Q_j^g(z_j^g \mid z_{j-1}^g) \right] \\
& \times \sum_{z_{m_g}^g} Q_{m_g}^g(z_{m_g}^g \mid z_{m_g-1}^g) \, Q_1^{g+1}(k \mid z_{m_g}^g).
\end{aligned}$$

To simplify the notation, we define $V^g_{m_g}(k \mid l) = Q^{g+1}_1(k \mid l)$, and, recursively for $j \in \{1, \ldots, m_g - 1\}$,

$$V^g_j(k \mid l) = \sum_{l'=1}^{K} Q^g_{j+1}(l' \mid l)\, V^g_{j+1}(k \mid l').$$

Then, we see that the $\mathcal{N}$ function can be reduced to:

$$\mathcal{N}_g(k) = \sum_{l=1}^{K} \frac{Q^g_1(l \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g \neq 1]}\, Q^g_1(l \mid z^{g-1}_{m_{g-1}})}{\mathcal{N}_{g-1}(l)} V^g_1(k \mid l).$$

It remains to be shown how to compute $V^g_1(k \mid l)$. We guess that $V^g_j(k \mid l)$ can be written as:

$$V^g_j(k \mid l) = v^g_{j,k} + u^g_j \mathbb{1}\left[k = l\right],$$

since $Q^{g+1}_1(k \mid l) = a^{g+1}_{1,k} + b^{g+1}_1 \mathbb{1}\left[k = l\right]$. This ansatz is clearly correct if $j = m_g$, by definition of $V^g_{m_g}(k \mid l)$, in which case: $v^g_{m_g,k} = a^{g+1}_{1,k}$ and $u^g_{m_g} = b^{g+1}_1$. When $j \in \{1, \ldots, m_g - 1\}$, backward induction shows that

$$\begin{aligned}
V^g_j(k' \mid k) &= \sum_{l=1}^{K} Q^g_{j+1}(l \mid k)\, V^g_{j+1}(k' \mid l) \\
&= \sum_{l=1}^{K} Q^g_{j+1}(l \mid k) \left(v^g_{j+1,k'} + u^g_{j+1} \mathbb{1}\left[k' = l\right]\right) \\
&= v^g_{j+1,k'} \sum_{l=1}^{K} Q^g_{j+1}(l \mid k) + u^g_{j+1} Q^g_{j+1}(k' \mid k) \\
&= v^g_{j+1,k'} \sum_{l=1}^{K} \left(a^g_{j+1,l} + b^g_{j+1} \mathbb{1}\left[l = k\right]\right) + u^g_{j+1} \left(a^g_{j+1,k'} + b^g_{j+1} \mathbb{1}\left[k' = k\right]\right) \\
&= v^g_{j+1,k'} \left(\sum_{l=1}^{K} a^g_{j+1,l} + b^g_{j+1}\right) + u^g_{j+1} a^g_{j+1,k'} + u^g_{j+1} b^g_{j+1} \mathbb{1}\left[k' = k\right]. \\
&= v^g_{j+1,k'} + u^g_{j+1} a^g_{j+1,k'} + u^g_{j+1} b^g_{j+1} \mathbb{1}\left[k' = k\right].
\end{aligned}$$

Therefore, our guess is correct for j, as long as it is for $j + 1$, since

$$v^g_{j,k'} = v^g_{j+1,k'} + u^g_{j+1} a^g_{j+1,k'}, \qquad u^g_j = u^g_{j+1} b^g_{j+1}.$$

□

Proof of Proposition 5. We start from

$$\mathbb{P}\left[\tilde{Z}^g = \tilde{z}^g \mid z^{-g}, \tilde{z}^{1:(g-1)}\right] \propto \frac{Q^g_1(\tilde{z}^g_1 \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g \neq 1]}\, Q^g_1(\tilde{z}^g_1 \mid z^{g-1}_{m_{g-1}})}{\mathcal{N}_{g-1}(z^g_1)} \left[\prod_{j=2}^{m_g} Q^g_j(\tilde{z}^g_j \mid \tilde{z}^g_{j-1})\right]$$
$$\times\, Q^{g+1}_1(z^{g+1}_1 \mid \tilde{z}^g_{m_g}),$$

and marginalize over the remaining $m_g - 1$ components, finding:

$$\mathbb{P}\left[\tilde{Z}_1^g = k \mid z^{-g}, \tilde{z}^{1:(g-1)}\right] \propto \frac{Q_1^g(k \mid \tilde{z}_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}\, Q_1^g(k \mid z_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(k)} V_1^g(z_1^{g+1} \mid k).$$

Above, V_j^g is as defined in the proof of Proposition 4. Given $\tilde{Z}_1^g$, we can sample each $\tilde{Z}_j^g$ sequentially, from $j = 2$ to $j = m_g$. It is easy to verify that, at the j-th step, we need to sample $\tilde{Z}_j^g$ from:

$$\mathbb{P}\left[\tilde{Z}_j^g = k \mid z^{-g}, \tilde{z}^{1:(g-1)}, \tilde{z}_{1:(j-1)}^g\right] \propto Q_j^g(k \mid \tilde{z}_{j-1}^g)\, V_j^g(z_1^{g+1} \mid k).$$

□

Proof of Proposition 6. Starting from

$$\mathcal{N}_g(\{k^a, k^b\}) = \sum_{\{l_1^a, l_1^b\}, \ldots, \{l_{m_g}^a, l_{m_g}^b\}} \frac{\bar{Q}_1^g(\{l_1^a, l_1^b\} \mid z_{m_{g-1}}^{g-1})^{\mathbb{1}[g\neq 1]}\, \bar{Q}_1^g(\{l_1^a, l_1^b\} \mid \tilde{z}_{m_{g-1}}^{g-1})}{\mathcal{N}_{g-1}(\{l_1^a, l_1^b\})}$$
$$\times \left[\prod_{j=2}^{m_g} \bar{Q}_j^g(\{l_j^a, l_j^b\} \mid \{l_{j-1}^a, l_{j-1}^b\})\right] \times \bar{Q}_1^{g+1}(\{k^a, k^b\} \mid \{l_{m_g}^a, l_{m_g}^b\}),$$

we proceed as in the proof of Proposition 4, defining:

$$V_{m_g}^g(\{k^a, k^b\} \mid \{l^a, l^b\}) = \bar{Q}_1^{g+1}(\{k^a, k^b\} \mid \{l^a, l^b\}).$$

Similarly, for $j = 1, \ldots, m_g - 1$, we recursively define:

$$V_j^g(\{k^a, k^b\} \mid \{l^a, l^b\}) = \sum_{\{h^a, h^b\}} \bar{Q}_{j+1}^g(\{h^a, h^b\} \mid \{l^a, l^b\})\, V_{j+1}^g(\{k^a, k^b\} \mid \{h^a, h^b\}).$$

Using Lemma 1, we can write:

$$V_{m_g}^g(\{k^a, k^b\} \mid \{l^a, l^b\}) = a_{1,k^a}^{g+1} a_{1,k^b}^{g+1}(2 - \delta_{k^a,k^b}) + (b_1^{g+1})^2 \delta_{\{k^a,k^b\},\{l^a,l^b\}}$$
$$+ b_1^{g+1} \frac{a_{1,k^a}^{g+1}\left(\delta_{k^b,l^a} + \delta_{k^b,l^b}\right) + a_{1,k^b}^{g+1}\left(\delta_{k^a,l^a} + \delta_{k^a,l^b}\right)}{1 + \delta_{k^a,k^b}}.$$

It is now time for an ansatz. For $j = 1, \ldots, m_g - 1$, we guess that:

$$V_j^g(\{k^a, k^b\} \mid \{l^a, l^b\}) = v_{j,\{k^a,k^b\}}^g(2 - \delta_{k^a,k^b}) + u_j^g \delta_{\{k^a,k^b\},\{l^a,l^b\}}$$
$$+ \frac{w_{j,k^a}^g\left(\delta_{k^b,l^a} + \delta_{k^b,l^b}\right) + w_{j,k^b}^g\left(\delta_{k^a,l^a} + \delta_{k^a,l^b}\right)}{1 + \delta_{k^a,k^b}}.$$

Assuming that our guess is correct for $j + 1$, with some $j \in \{1, \ldots, m_g - 1\}$, we can write:

$$V_j^g(\{k^a, k^b\} \mid \{l^a, l^b\}) = \sum_{\{h^a, h^b\}} \bar{Q}_{j+1}^g(\{h^a, h^b\} \mid \{l^a, l^b\})\, V_{j+1}^g(\{k^a, k^b\} \mid \{h^a, h^b\})$$

$$
\begin{aligned}
&= v^g_{j+1,\{k^a,k^b\}}(2-\delta_{k^a,k^b}) \sum_{\{h^a,h^b\}} \bar{Q}^g_{j+1}(\{h^a,h^b\} \mid \{l^a,l^b\}) \\
&\quad + \frac{w^g_{j+1,k^a}}{1+\delta_{k^a,k^b}} \sum_{\{h^a,h^b\}} \left(\delta_{k^b,h^a}+\delta_{k^b,h^b}\right) \bar{Q}^g_{j+1}(\{h^a,h^b\} \mid \{l^a,l^b\}) \\
&\quad + \frac{w^g_{j+1,k^b}}{1+\delta_{k^a,k^b}} \sum_{\{h^a,h^b\}} \left(\delta_{k^a,h^a}+\delta_{k^a,h^b}\right) \bar{Q}^g_{j+1}(\{h^a,h^b\} \mid \{l^a,l^b\}) \\
&\quad + u^g_{j+1}\, \bar{Q}^g_{j+1}(\{k^a,k^b\} \mid \{l^a,l^b\}) \\
&= v^g_{j+1,\{k^a,k^b\}}(2-\delta_{k^a,k^b}) + u^g_{j+1}\, \bar{Q}^g_{j+1}(\{k^a,k^b\} \mid \{l^a,l^b\}) \\
&\quad + \frac{1}{1+\delta_{k^a,k^b}} \left[w^g_{j+1,k^a}\Gamma(k^b \mid \{l^a,l^b\}) + w^g_{j+1,k^b}\Gamma(k^a \mid \{l^a,l^b\}) \right].
\end{aligned}
$$

Above, we have defined:

$$
\Gamma(k \mid \{l^a,l^b\}) = \sum_{\{h^a,h^b\}} \left(\delta_{k,h^a}+\delta_{k,h^b}\right) \bar{Q}^g_{j+1}(\{h^a,h^b\} \mid \{l^a,l^b\}).
$$

This can be further simplified:

$$
\begin{aligned}
\Gamma(k \mid \{l^a,l^b\}) &= \sum_{\{h^a,h^b\}} \left(\delta_{k,h^a}+\delta_{k,h^b}\right) \bar{Q}^g_{j+1}(\{h^a,h^b\} \mid \{l^a,l^b\}) \\
&= 2\sum_{h} \delta_{k,h}\bar{Q}^g_{j+1}(\{h,h\} \mid \{l^a,l^b\}) + \sum_{\{h^a,h^b\},h^a\neq h^b} \left(\delta_{k,h^a}+\delta_{k,h^b}\right) \bar{Q}^g_{j+1}(\{h^a,h^b\} \mid \{l^a,l^b\}) \\
&= 2\bar{Q}^g_{j+1}(\{k,k\} \mid \{l^a,l^b\}) + \sum_{\{h^a,h^b\},h^a\neq h^b} \left(\delta_{k,h^a}+\delta_{k,h^b}\right) \bar{Q}^g_{j+1}(\{h^a,h^b\} \mid \{l^a,l^b\}) \\
&= 2\bar{Q}^g_{j+1}(\{k,k\} \mid \{l^a,l^b\}) + \frac{1}{2} \sum_{h^a,h^b,h^a\neq h^b} \left(\delta_{k,h^a}+\delta_{k,h^b}\right) \bar{Q}^g_{j+1}(\{h^a,h^b\} \mid \{l^a,l^b\}) \\
&= 2\bar{Q}^g_{j+1}(\{k,k\} \mid \{l^a,l^b\}) + \frac{1}{2} \sum_{h^a,h^b,h^a\neq h^b} \left(\delta_{k,h^a}+\delta_{k,h^b}\right) Q^g_{j+1}(h^a \mid l^a) Q^g_{j+1}(h^b \mid l^b) \\
&\quad + \frac{1}{2} \sum_{h^a,h^b,h^a\neq h^b} \left(\delta_{k,h^a}+\delta_{k,h^b}\right) Q^g_{j+1}(h^a \mid l^b) Q^g_{j+1}(h^b \mid l^a) \\
&= 2\bar{Q}^g_{j+1}(\{k,k\} \mid \{l^a,l^b\}) + \frac{1}{2}Q^g_{j+1}(k \mid l^a) \sum_{h^b\neq k} Q^g_{j+1}(h^b \mid l^b) \\
&\quad + \frac{1}{2}Q^g_{j+1}(k \mid l^b) \sum_{h^a\neq k} Q^g_{j+1}(h^a \mid l^a) + \frac{1}{2}Q^g_{j+1}(k \mid l^b) \sum_{h^b\neq k} Q^g_{j+1}(h^b \mid l^a) \\
&\quad + \frac{1}{2}Q^g_{j+1}(k \mid l^a) \sum_{h^a\neq k} Q^g_{j+1}(h^a \mid l^b) \\
&= 2\bar{Q}^g_{j+1}(\{k,k\} \mid \{l^a,l^b\}) \\
&\quad + \frac{1}{2}Q^g_{j+1}(k \mid l^a)\left[1-Q^g_{j+1}(k \mid l^b)\right] + \frac{1}{2}Q^g_{j+1}(k \mid l^b)\left[1-Q^g_{j+1}(k \mid l^a)\right] \\
&\quad + \frac{1}{2}Q^g_{j+1}(k \mid l^b)\left[1-Q^g_{j+1}(k \mid l^a)\right] + \frac{1}{2}Q^g_{j+1}(k \mid l^a)\left[1-Q^g_{j+1}(k \mid l^b)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= 2Q^g_{j+1}(k \mid l^a)Q^g_{j+1}(k \mid l^b) + Q^g_{j+1}(k \mid l^a)\left[1 - Q^g_{j+1}(k \mid l^b)\right] \\
&\qquad + Q^g_{j+1}(k \mid l^b)\left[1 - Q^g_{j+1}(k \mid l^a)\right] \\
&= Q^g_{j+1}(k \mid l^a) + Q^g_{j+1}(k \mid l^b) \\
&= 2a^g_{j+1,k} + b^g_{j+1}\left(\delta_{k,l^a} + \delta_{k,l^b}\right).
\end{aligned}
$$

Therefore, we can rewrite V^g_j, as follows:

$$
\begin{aligned}
&V^g_j(\{k^a, k^b\} \mid \{l^a, l^b\}) \\
&\quad = v^g_{j+1,\{k^a,k^b\}}(2 - \delta_{k^a,k^b}) + u^g_{j+1}\,\bar{Q}^g_{j+1}(\{k^a,k^b\} \mid \{l^a,l^b\}) \\
&\qquad + \frac{w^g_{j+1,k^a}\Gamma(k^b \mid \{l^a,l^b\}) + w^g_{j+1,k^b}\Gamma(k^a \mid \{l^a,l^b\})}{1 + \delta_{k^a,k^b}} \\
&\quad = v^g_{j+1,\{k^a,k^b\}}(2 - \delta_{k^a,k^b}) \\
&\qquad + \frac{w^g_{j+1,k^a}\left[2a^g_{j+1,k^b} + b^g_{j+1}\left(\delta_{k^b,l^a} + \delta_{k^b,l^b}\right)\right] + w^g_{j+1,k^b}\left[2a^g_{j+1,k^a} + b^g_{j+1}\left(\delta_{k^a,l^a} + \delta_{k^a,l^b}\right)\right]}{1 + \delta_{k^a,k^b}} \\
&\qquad + u^g_{j+1}\Bigg[a_{j+1,k^a}a_{j+1,k^b}(2 - \delta_{k^a,k^b}) + (b^g_{j+1})^2\delta_{\{l^a,l^b\},\{k^a,k^b\}} \\
&\qquad\qquad + b^g_{j+1}\frac{a^g_{j+1,k^a}\left(\delta_{k^b,l^a} + \delta_{k^b,l^b}\right) + a^g_{j+1,k^b}\left(\delta_{k^a,l^a} + \delta_{k^a,l^b}\right)}{1 + \delta_{k^a,k^b}}\Bigg].
\end{aligned}
$$

Now, we only need to collect these terms to obtain the recursion rules:

$$
\begin{aligned}
u^g_j &= u^g_{j+1}(b^g_{j+1})^2, \\
v^g_{j,\{k^a,k^b\}} &= v^g_{j+1,\{k^a,k^b\}} + u^g_{j+1}a^g_{j+1,k^a}a^g_{j+1,k^b} + w^g_{j+1,k^a}a^g_{j+1,k^b} + w^g_{j+1,k^b}a^g_{j+1,k^a}, \\
w^g_{j,k} &= w^g_{j+1,k}b^g_{j+1} + u^g_{j+1}a^g_{j+1,k}b^g_{j+1}.
\end{aligned}
$$

Finally, the $\mathcal{N}$ function for the g-th group is given by:

$$
\begin{aligned}
\mathcal{N}_g(\{k^a,k^b\}) &= \sum_{\{l^a,l^b\}} \frac{\bar{Q}^g_1(\{l^a,l^b\} \mid z^{g-1}_{m_{g-1}})\,\bar{Q}^g_1(\{l^a,l^b\} \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g\neq 1]}}{\mathcal{N}_{g-1}(\{l^a,l^b\})} V^g_1(\{k^a,k^b\} \mid \{l^a,l^b\}) \\
&= v^g_{1,\{k^a,k^b\}}(2 - \delta_{k^a,k^b}) \sum_{\{l^a,l^b\}} \frac{\bar{Q}^g_1(\{l^a,l^b\} \mid z^{g-1}_{m_{g-1}})\,\bar{Q}^g_1(\{l^a,l^b\} \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g\neq 1]}}{\mathcal{N}_{g-1}(\{l^a,l^b\})} \\
&\quad + \frac{w^g_{1,k^a}}{1 + \delta_{k^a,k^b}} \sum_{\{l^a,l^b\}} \frac{\bar{Q}^g_1(\{l^a,l^b\} \mid z^{g-1}_{m_{g-1}})\,\bar{Q}^g_1(\{l^a,l^b\} \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g\neq 1]}}{\mathcal{N}_{g-1}(\{l^a,l^b\})}\left(\delta_{k^b,l^a} + \delta_{k^b,l^b}\right) \\
&\quad + \frac{w^g_{1,k^b}}{1 + \delta_{k^a,k^b}} \sum_{\{l^a,l^b\}} \frac{\bar{Q}^g_1(\{l^a,l^b\} \mid z^{g-1}_{m_{g-1}})\,\bar{Q}^g_1(\{l^a,l^b\} \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g\neq 1]}}{\mathcal{N}_{g-1}(\{l^a,l^b\})}\left(\delta_{k^a,l^a} + \delta_{k^a,l^b}\right) \\
&\quad + u^g_1\frac{\bar{Q}^g_1(\{k^a,k^b\} \mid z^{g-1}_{m_{g-1}})\,\bar{Q}^g_1(\{k^a,k^b\} \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g\neq 1]}}{\mathcal{N}_{g-1}(\{k^a,k^b\})}.
\end{aligned}
$$

Let us now define

$$C^g(\{l^a, l^b\}) = \frac{\bar{Q}^g_1(\{l^a, l^b\} \mid z^{g-1}_{m_{g-1}})\, \bar{Q}^g_1(\{l^a, l^b\} \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g\neq 1]}}{\mathcal{N}_{g-1}(\{l^a, l^b\})},$$

and

$$D^g(k) = C^g(\{k, k\}) + \sum_{l=1}^{K} C^g(\{k, l\}).$$

Then, we can write the $\mathcal{N}$ function more compactly, as follows:

$$\begin{aligned}
\mathcal{N}_g(\{k^a, k^b\}) &= v^g_{1,\{k^a,k^b\}}(2 - \delta_{k^a,k^b}) \sum_{\{l^a,l^b\}} C^g(\{l^a, l^b\}) + u^g_1\, C^g(\{k^a, k^b\}) \\
&\quad + \frac{w^g_{1,k^a}}{1 + \delta_{k^a,k^b}} \sum_{\{l^a,l^b\}} C^g(\{l^a, l^b\}) \left(\delta_{k^b,l^a} + \delta_{k^b,l^b}\right) \\
&\quad + \frac{w^g_{1,k^b}}{1 + \delta_{k^a,k^b}} \sum_{\{l^a,l^b\}} C^g(\{l^a, l^b\}) \left(\delta_{k^a,l^a} + \delta_{k^a,l^b}\right) \\
&= (2 - \delta_{k^a,k^b}) v^g_{1,\{k^a,k^b\}} \sum_{\{l^a,l^b\}} C^g(\{l^a, l^b\}) + u^g_1\, C^g(\{k^a, k^b\}) \\
&\quad + \frac{w^g_{1,k^a} D^g(k^b) + w^g_{1,k^b} D^g(k^a)}{1 + \delta_{k^a,k^b}}.
\end{aligned}$$

□

Proof of Proposition 7. We proceed as in the proof of Proposition 5, using the notation developed in the proof of Proposition 6. By marginalizing over the remaining $m_g - 1$ components, it turns out that the marginal distribution of $\tilde{Z}^g_1$ is:

$$\mathbb{P}\left[\tilde{Z}^g_1 = \{l^a, k^b\} \mid z^{-g}, \tilde{z}^{1:(g-1)}\right] \propto \frac{\bar{Q}^g_1(\{k^a, k^b\} \mid z^{g-1}_{m_{g-1}})\, \bar{Q}^g_1(\{k^a, k^b\} \mid \tilde{z}^{g-1}_{m_{g-1}})^{\mathbb{1}[g\neq 1]}}{\mathcal{N}_{g-1}(\{k^a, k^b\})} V^g_1(z^{g+1}_1 \mid \{k^a, k^b\}).$$

Since we have already obtained $V^g_1(z^{g+1}_1 \mid \{k^a, k^b\})$, by computing the $\mathcal{N}$ function, sampling $\tilde{Z}^g_1$ can be performed easily, in $\mathcal{O}(K^2)$ time. Given $\tilde{Z}^g_1$, we proceed to sample each $\tilde{Z}^g_j$ sequentially, from $j = 2$ to $j = m_g$. It is easy to verify that, at the j-th step, we sample $\tilde{Z}^g_j$ from:

$$\mathbb{P}\left[\tilde{Z}^g_j = \{k^a, k^b\} \mid z^{-g}, \tilde{z}^{1:(g-1)}\right] \propto \bar{Q}^g_j(\{k^a, k^b\} \mid \tilde{z}^g_{j-1})\, V^g_j(z^{g+1}_1 \mid \{k^a, k^b\}).$$

Again, this can be easily performed in $\mathcal{O}(K^2)$ time, for each j. □

Proof of Proposition 8. The result for $F_1(k)$ is immediate from (S7). For $j \in \{2, \ldots, p\}$,

$$\frac{F_{j+1}(k)}{f_{j+1}(x_{j+1} \mid k)} = \sum_{l=1}^{K} Q_{j+1}(k \mid l)\, F_j(l) = a_{j+1,k} \sum_{l=1}^{K} F_j(l) + b_{j+1} F_j(k).$$

□

Proof of Proposition 9. The case if $F_1(\{k^a, k^b\})$ follows directly from (S8). For $j \in \{2, \ldots, p\}$,

$$\begin{aligned}
\frac{F_{j+1}(\{k^a,k^b\})}{f_{j+1}(z_{j+1} \mid \{k^a,k^b\})} &= \sum_{\{l^a,l^b\}} \bar{Q}_{j+1}(\{k^a,k^b\} \mid \{l^a,l^b\})\, F_j(\{l^a,l^b\}) \\
&= a_{j+1,k^a} a_{j+1,k^b}(2-\delta_{k^a,k^b}) \sum_{\{l^a,l^b\}} F_j(\{l^a,l^b\}) + (b_{j+1})^2 F_j(\{k^a,k^b\}) \\
&\quad + b_{j+1}\frac{a_{j+1,k^a}}{1+\delta_{k^a,k^b}} \sum_{\{l^a,l^b\}} \left(\delta_{k^b,l^a}+\delta_{k^b,l^b}\right) F_j(\{l^a,l^b\}) \\
&\quad + b_{j+1}\frac{a_{j+1,k^b}}{1+\delta_{k^a,k^b}} \sum_{\{l^a,l^b\}} \left(\delta_{k^a,l^a}+\delta_{k^a,l^b}\right) F_j(\{l^a,l^b\}) \\
&= a_{j+1,k^a} a_{j+1,k^b}(2-\delta_{k^a,k^b}) \sum_{\{l^a,l^b\}} F_j(\{l^a,l^b\}) + (b_{j+1})^2 F_j(\{k^a,k^b\}) \\
&\quad + b_{j+1}\frac{a_{j+1,k^a}}{1+\delta_{k^a,k^b}} \left(F_j(\{k^b,k^b\}) + \sum_{l=1}^{K} F_j(\{l,k^b\}) \right) \\
&\quad + b_{j+1}\frac{a_{j+1,k^b}}{1+\delta_{k^a,k^b}} \left(F_j(\{k^a,k^a\}) + \sum_{l=1}^{K} F_j(\{l,k^a\}) \right).
\end{aligned}$$

□

[1] Katsevich, E. & Sabatti, C. Multilayer knockoff filter: controlled variable selection at multiple resolutions. *Ann. Appl. Stat.* **13**, 1–33 (2019).

[2] Sesia, M., Sabatti, C. & Candès, E. J. Gene hunting with hidden markov model knockoffs. *Biometrika* **106**, 1–18 (2019).

[3] Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

[4] Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).

[5] Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

[6] O'Connell, J. *et al.* Haplotype estimation for biobank scale datasets. *Nat. Genet.* **48**, 817–820 (2016).

[7] Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R, packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).

[8] Abraham, G., Kowalczyk, A., Zobel, J. & Inouye, M. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genet. Epidemiol.* **37**, 184–195 (2013).

[9] Vilhjá, B. J., lmsson *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

[10] Wei, Z. *et al.* Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* **92**, 1008–1012 (2013).

[11] Botta, V., Louppe, G., Geurts, P. & Wehenkel, L. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS One* **9** (2014).

[12] Bellot, P. & Pérez-Enciso, M. Can deep learning improve genomic prediction of complex human traits? *Genetics* **210**, 809–819 (2018).

[13] Klasen, J. R. *et al.* A multi-marker association method for genome-wide association studies without the need for population structure correction. *Nat. Commun.* **7** (2016).

[14] Zeng, Y. & Breheny, P. The biglasso package: A memory-and computation-efficient solver for lasso model fitting with big data in R. *arXiv* (2017).

[15] Candès, E. J., Fan, Y., Janson, L. & Lv, J. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *J. R. Stat. Soc. B.* **80**, 551–577 (2018).

[16] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.* **57**, 289–300 (1995).

[17] Simes, R. J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754 (1986).

[18] Katsevich, E., Sabatti, C. & Bogomolov, M. Controlling FDR while highlighting distinct discoveries. *arXiv* (2018).

[19] Siegmund, D. O., Zhang, N. R. & Yakir, B. False discovery rate for scanning statistics. *Biometrika* **98**, 979–985 (2011).

[20] Brzyski, D. *et al.* Controlling the rate of GWAS false discoveries. *Genetics* **205**, 61–75 (2017).

[21] Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).

[22] Wang, G., Sarkar, A. K., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv* (2018).

[23] Barber, R. F. & Candès, E. J. Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085 (2015).

[24] Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Cambridge University Press, 2010).

[25] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

[26] McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495 (2010).

[27] Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research* **28**, 352–355 (2000).